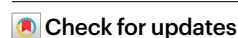


Single-round evolution of RNA aptamers with GRAPE-LM

Received: 3 April 2025

Accepted: 7 January 2026

Published online: 06 February 2026



Jun Zhang^{1,2,10}✉, Ju Zhang^{3,4,10}, Shaoxuan Tang^{1,2}, Chuancheng Liu^{3,4}, Yushan Cai^{3,4}, Hao Zeng^{1,2}, Xiangjie Meng^{5,6}, Bei Liu⁴, Yang Zhang^{5,6,7,11}✉ & Yu Wang^{3,4,8,9,11}✉

The directed evolution of biomolecules is an iterative process. Although advancements in language models have expedited protein evolution, effectively evolving RNA remains a challenge. RNA aptamers, selected for their binding properties, provide an ideal system to address this challenge, yet traditional aptamer discovery still relies on labor-intensive, multi-round screening. Here we introduce GRAPE-LM (generator of RNA aptamers powered by activity-guided evolution and language model), a generative artificial intelligence framework designed for the one-round evolution of RNA aptamers. GRAPE-LM integrates a transformer-based conditional autoencoder with nucleic acid language models and is guided by CRISPR–Cas-based aptamer screening data derived from intracellular environments. We validate GRAPE-LM on three disparate targets: the human T cell receptor CD3ε, the receptor-binding domain of the SARS-CoV-2 spike protein and the human oncogenic transcription factor c-Myc (an intracellular disordered protein). GRAPE-LM, informed with only a single round of CRISPR–Cas-based screening, successfully obtains RNA aptamers that outperform those driven from multiple rounds of human selection and optimization.

The directed evolution of biomolecules, particularly RNAs and proteins, plays a pivotal role in the development of novel therapeutics, biosensors and other functional biomolecules¹. Integrating computational methods, especially artificial intelligence (AI), can enhance these processes^{2–4}. For example, EVOLVEpro⁵ has demonstrated that protein evolution can be accelerated through the integration of protein language models (LMs) trained on experimental data. However, a parallel exploration of RNA evolution remains largely untapped, particularly in areas beyond its predictable coding capacity, such as its complex and less understood binding interactions.

RNA aptamers⁶, defined by binding affinity to specific targets, provide a model for studying the evolution of RNAs for binding beyond predictable coding^{7–9}. Traditionally, aptamer selection relies on systematic evolution of ligands by exponential enrichment (SELEX)¹⁰. This procedure is highly manual and labor intensive. Moreover, the selection process is conducted in solution, thus lacking biological relevance. These challenges underscore the need for aptamer evolution methods to select aptamers with functionality and efficacy in cellular contexts. CRISmers (CRISPR–Cas-based aptamer screening system), an intracellular system, was recently developed in our laboratory¹¹. This

¹School of Artificial Intelligence, Shenzhen University, Shenzhen, China. ²National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China. ³State Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. ⁴College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China. ⁵School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China. ⁶School of Biomedical Engineering, Harbin Institute of Technology, Shenzhen, China. ⁷Shenzhen Loop Area Institute, Shenzhen, China. ⁸Shenzhen Key Laboratory of Genome Manipulation and Biosynthesis, Shenzhen, China. ⁹Shanghai Key Laboratory of Rare Disease Gene Editing and Cell Therapy, Fudan University, Shanghai, China. ¹⁰These authors contributed equally: Jun Zhang, Ju Zhang. ¹¹These authors jointly supervised this work: Yu Wang, Yang Zhang. ✉e-mail: junzhang@szu.edu.cn; yang.zhang2020@hotmail.com; yuwang@post.harvard.edu

approach accounts for factors such as intracellular competition, folding dynamics and interaction specificity, offering a more biologically relevant context for aptamer discovery. An independent study from Su-Tobon et al.¹² echoed the power of intracellular aptamer screening using the CRISPR–Cas system. However, these aptamer selection methods still rely on tedious rounds of human operation, resulting in low throughput. Additionally, the aptamer selection library of CRISmers is relatively small due to limitations in transfection efficiency and the limited capacity of cells to uptake the sequences, potentially missing a large population of relevant sequences.

Computation-based approaches have facilitated the process of aptamer evolution. For example, RaptGen¹³ and AptADiff¹⁴ combine generative AI models with SELEX data to generate new active aptamers. RhoDesign¹⁵ leverages structural data to guide the generation of RNA aptamers through a transformer-based framework of reverse folding^{16–18}. AptA-MCTS¹⁹, AptATrans²⁰ and AptABLE²¹ integrate an aptamer–protein interaction prediction model and reinforcement learning with Monte Carlo tree search (MCTS) to actively explore promising aptamer sequence spaces informed by target proteins. Together, these methods underscore the potential of AI to enable *de novo* design of functional aptamers through computational prediction. However, current AI-driven approaches for aptamer generation still face notable limitations. For instance, although RaptGen and AptADiff employ sequence patterns to map aptamer characteristics, they cannot independently produce high-activity candidates. This method requires supplementary computational tools—such as DeepBind²² for binding affinity analysis and Bayesian optimization algorithms—to filter and refine its outputs. RhoDesign relies on preexisting aptamer structures as templates to guide new designs, which limits its broader application due to sparse availability of RNA structure data^{2,23}. Consequently, there is a need to explore alternative approaches to mitigate these constraints.

Here we present GRAPE-LM, a generative framework that incorporates the nucleic acid LM and a transformer-based conditional autoencoder (CAE) to evolve RNA aptamers. GRAPE-LM uses next-generation sequencing (NGS) enrichment data from CRISmers, ensuring biologically relevant outputs. The framework surpasses variational

autoencoder (VAE)¹³, masked language modeling²⁴ and diffusion²⁵ in generating diverse or active aptamer sequences. Notably, a one-shot iteration using GRAPE-LM informed by a single round of CRISmers successfully evolved RNA aptamers that exhibit higher binding activities than previously published aptamers from the traditional methods of multiple-round screening and optimization. This achievement was demonstrated with three disparate protein targets: the human T cell receptor CD3ε (hereafter ‘CD3ε’), the SARS-CoV-2 spike receptor-binding domain (RBD) (hereafter ‘RBD’) and the human c-Myc transcription factor (an intracellular disordered protein). GRAPE-LM requires no prior knowledge of aptamers and can initiate RNA evolution using a relatively small starting library (approximately 2×10^8), using AI’s generative power to produce numerous functional sequence candidates. This contrasts with traditional SELEX methods that require a large initial library (normally $1 \times 10^{14–16}$) to achieve meaningful results.

Results

Framework of GRAPE-LM

Different from VAE, diffusion or reverse folding frameworks^{13–15}, GRAPE-LM uses CAE (Fig. 1a), a more straightforward and effective deep-learning-based framework (Extended Data Fig. 1a), to evolve RNA aptamers guided by quantitative NGS enrichment data (Supplementary Table 1) from CRISmers (Extended Data Fig. 1b). The CAE outperforms VAE and masked language modeling frameworks in terms of generated diversity (Fig. 1b). Although the diffusion framework generates diverse sequences effectively, its inherent denoising training scheme prevents incorporating activity guidance modules. This limitation is evidenced by poor performance in Pearson’s correlation coefficient (Fig. 1c).

Incorporation of LMs. To validate the impact of LM integration in GRAPE-LM, we conducted ablation experiments comparing eight configurations: two non-LM controls (long short-term memory (LSTM) network encoder and convolutional neural network (CNN) encoder, with one-hot coding) and six variants using distinct nucleic acid LMs, namely RNA-FM²⁶, RNA-BERT²⁷, RNAErnie²⁸, EvoFlow-RNA²⁹, RiNALMo³⁰ (see Supplementary Fig. 1a for the network architectures of these RNA LMs) and Evo³¹ (see Supplementary Fig. 1b for its network architecture).

Fig. 1 | Illustration of GRAPE-LM and computational results. **a**, Overall framework of GRAPE-LM, a novel generative framework using a CAE powered by a nucleic acid LM to evolve RNA aptamers. GRAPE-LM comprises three components: LM-powered encoder, which embeds RNA sequences into a continuous latent semantic space; activity guidance module, which constrains the latent space to organize the RNA sequences according to their pseudo-activity levels, enabling targeted sampling; and transformer-based decoder, which generates novel RNA aptamer candidates by sampling from potential high-activity regions of the latent space. During training, the backbone LM is frozen (snowflake symbol), and only the adapter is fine-tuned. Input data were generated by CRISmers¹¹ unless otherwise noted. Icons in figures were created with BioRender.com. **b**, Comparison of generative frameworks for CD3ε and RBD. The generated diversity was calculated based on the samples generated by a model and the seed sequences. **c**, Comparison of CAE and diffusion framework in terms of activity guidance. The effectiveness of activity guidance was measured by Pearson’s correlation coefficient (Pearson’s *r*) between the predicted pseudo-activity score and the ground truth derived from CRISmers. **d**, The ablation study of the nucleic acid LMs on GRAPE-LM. The recovery rate was calculated using a model on the test set ($n = 6,499$ samples for CD3ε, $n = 7,740$ samples for RBD). The statistical analysis was conducted by a two-sided Student’s *t*-test using RNA-FM as the baseline (**** $P < 0.0001$). **e, f**, Comparison of GRAPE-LM with different nucleic acid LMs in terms of generated diversity and Pearson’s correlation coefficients in pseudo-activity prediction. **g**, The average time required for GRAPE-LM with different LMs to generate a 20-length RNA sequence with the same GPU resources. **h**, Evaluation of GRAPE-LM activity guidance. The precision-recall curve of predicted pseudo-activity score on the test sets illustrates the model’s hit rate under different readout thresholds. The random control refers to a

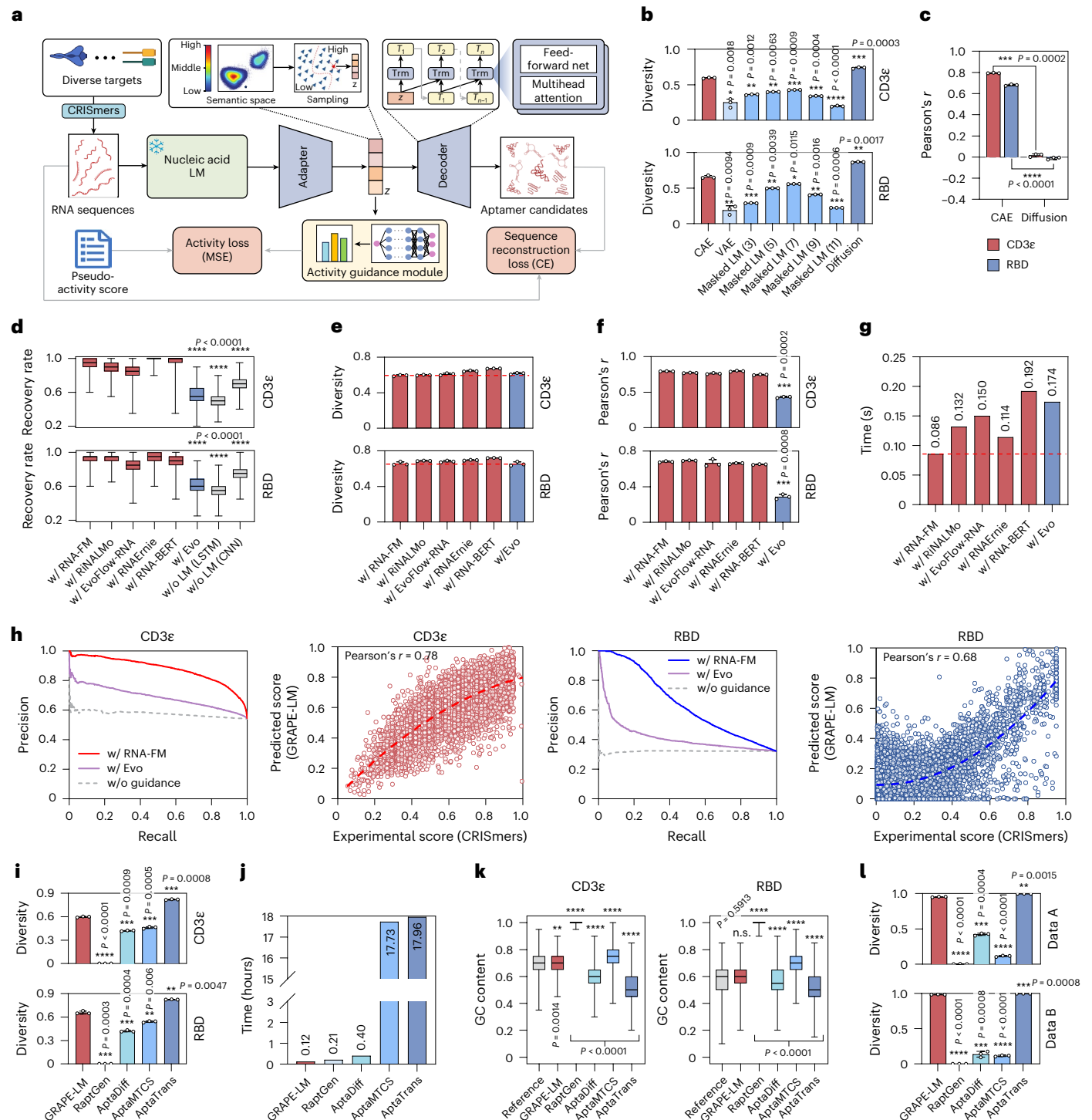
model without activity guidance training ($w = 0$). The scatter plot shows the effectiveness of GRAPE-LM’s activity guidance. The Pearson’s correlation coefficient is used as an evaluation index for the effectiveness of GRAPE-LM activity guidance. **i**, Comparison of generated diversity between GRAPE-LM and other methods. **j**, The time required to generate 5,000 RNA sequences of length 20 using different methods on the same device. **k**, Comparison of generation rationality between GRAPE-LM and other methods via GC content ($n = 5,000$ generated samples). The top 2,000 leads from pre-AI as a reference. Statistical significance was assessed by unpaired Welch’s two-sided *t*-test (**** $P < 0.0001$; ** $P < 0.01$). **l**, Comparative performance of GRAPE-LM and other methods on datasets from SELEX targeting human transglutaminase 2 (Data A) and human integrin αvβ3 (Data B) used by the creators of RaptGen¹³. The rectangular boxes in the box plot (**d** and **k**) represent the range of quartiles of the data, where the center line represents the median and the upper and lower edges represent the upper and lower quartiles, respectively. The line above the rectangular box represents the maximum value of the data; the line below represents the minimum value of the data. Data plot mean with s.d. ($n = 3$ experimental replicates (**b**, **c**, **e**, **f**, **i** and **l**)). The statistical analyses were conducted using a two-sided Student’s *t*-test (**** $P < 0.0001$; *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$). All runtime evaluations were under identical hardware and the same settings with a computing node equipped: one AMD EPYC 7642 CPU (512 GB RAM memory) and one NVIDIA RTX 4090 GPU (24 GB). Sampling and generation were performed using the same batch size (per-sequence rollout) and settings (one thread) across all models. No model-specific acceleration or customized inference optimization was applied beyond the standard PyTorch implementation. We reproduced the competing models strictly following the authors’ publicly released codes.

The results on CD3 ϵ and RBD show that integrating nucleic acid LMs into GRAPE-LM improves the recovery rate of RNA aptamer generation compared to non-LM approaches (Fig. 1d). This demonstrates that latent spaces powered by a pretrained LM, especially RNA LMs, better capture functional sequence features.

The non-LM approach, CNN-based encoder, achieved reasonable performance and even outperformed Evo in terms of recovery rate (Fig. 1d). This is likely because CNNs are sparse and computationally efficient architectures that excel at capturing local sequence patterns, enabling them to learn useful representations directly from sequence data. However, GRAPE-LM variants equipped with RNA LMs produced better results and consistently outperformed both non-LM

controls and Evo, reflecting the advantage of pretrained contextual representations in capturing global dependencies and functional features in RNA sequences. Notably, all tested RNA LMs achieved similarly good recovery rate and diversity for both target proteins (Fig. 1d,e). When evaluated using the Pearson's correlation coefficient (Pearson's r) between predicted activity scores and pseudo-activity scores, the RNA LMs demonstrated superior performance over the Evo model (Fig. 1f).

Although a diverse array of RNA LMs is equally viable (Fig. 1d–f), we selected RNA-FM due to its high computational efficiency on the same hardware and settings (Fig. 1g), considering the high-throughput requirements of the generative task.



Integration with activity guidance. Unlike conventional sequence similarity-based latent space¹³ (Extended Data Fig. 1a), we introduced an activity guidance module between the encoder and decoder to constrain the latent space (Extended Data Fig. 1b), enabling it to organize the RNA sequences according to their pseudo-activity levels. To evaluate the efficacy of the activity guidance strategy, we constructed a control model of GRAPE-LM by ablating the activity guidance (without guidance) while retaining all framework components. Both representative GRAPE-LM (with RNA-FM or Evo) and the control model were tested on two targets: CD3 ϵ and RBD (Fig. 1h). Removing the activity guidance led to the model losing its ability to distinguish between relatively high and low activity and degenerating into a random guessing level. By contrast, the complete GRAPE-LM model can determine the pseudo-activity level of sequences and achieve pseudo-activity prediction consistent with CRISmers screening on the test set (Pearson's correlation coefficient greater than 0.5; Fig. 1h). These results indicate that the activity guidance strategy can structure the latent space and enhance the model's ability to encode aptamer functional features.

To directly investigate the role that activity guidance may play in defining the latent space, we further evaluated different activity guidance loss weights and visualized the resulting latent space via t-distributed stochastic neighbor embedding (t-SNE) applied to GRAPE-LM's encoder outputs. The visualization of the latent vectors (Extended Data Fig. 2) revealed that increasing the activity guidance weight enhanced separation between high-ranking and low-ranking activity samples while driving generated sequences toward top active clusters—the latent spaces of the models without guidance ($w = 0$) or insufficient guidance cannot achieve this. Notably, complete reliance on activity guidance ($w = 1$, eliminating sequence reconstruction) caused generated sequences to diverge from natural distribution patterns despite maintaining activity discrimination. These results validate the critical role of the activity guidance strategy in encoding activity-relevant patterns within the latent space and emphasize the necessity of the CAE framework's dual optimization objectives: sequence generation and pseudo-activity guidance.

Comparison with other state-of-the-art methods for aptamer generation

To benchmark the performance of GRAPE-LM, we compared it (RNA-FM as the representative) to four existing methods, including RaptGen¹³, AptaDiff⁴, Apta-MCTS¹⁹ and AptaTrans²⁰. We evaluated and compared different methods on CD3 ϵ and RBD targets. Because Apta-MCTS and AptaTrans depend directly on target protein sequences, whereas AptaDiff functions via iterative denoising, the recovery rate metric cannot be reliably applied to any of these methods. Therefore, only RaptGen and GRAPE-LM were evaluated by the recovery rate, and GRAPE-LM consistently outperformed RaptGen (Extended Data Fig. 3a). In addition, GRAPE-LM outperformed three competing methods (RaptGen, AptaDiff, and Apta-MCTS) in terms of generated diversity (Fig. 1i). AptaTrans incorporates a built-in selection strategy in its MCTS process that filters out highly similar sequences, which results in greater diversity compared to GRAPE-LM. However, this improvement is accompanied by substantially higher computational cost (Fig. 1j). On identical hardware with the same settings, GRAPE-LM generates 5,000 samples in just 0.12 hours, whereas AptaTrans requires 17.96 hours. Considering the high-throughput demands of aptamer evolution, GRAPE-LM provides a complementary advantage in computational efficiency. Moreover, in terms of rationality of generated sequences, GRAPE-LM surpasses all four baseline methods by generating sequences with substantially lower GC content deviation from reference sequences (Fig. 1k).

For an objective comparison, we further evaluated different methods on the datasets used by RaptGen, the SELEX data targeting human transglutaminase 2 (Data A of Fig. 1l) and human integrin $\alpha\beta 3$ (Data B of Fig. 1l), respectively^{13,32}. We divided each dataset into training

and testing sets in an 8:2 ratio. Then, we retrained GRAPE-LM with the training set and evaluated it comparatively with other methods on the test set. The results are consistent with those of CD3 ϵ and RBD targets (Fig. 1i and Extended Data Fig. 3b). These findings also highlight that GRAPE-LM can be trained using data from conventional SELEX for aptamer evolution, demonstrating its generalizability.

Evolution of RNA aptamers with GRAPE-LM

To evaluate the capacity of GRAPE-LM to evolve RNA aptamers, we performed a pooled quantitative assessment based on CRISmers for sequence output libraries generated from GRAPE-LM (with RNA-FM or Evo) on two targets, CD3 ϵ and RBD, respectively (Fig. 2a). To mitigate redundancy and experimental bias, we applied Cd-hit³³ with an 80% similarity threshold to establish non-redundant sequence libraries (Supplementary Table 2) from cost-efficient evaluations of 10,000 (with RNA-FM) and 10,000 (with Evo) generated sequences.

Overall, CRISmers examination revealed that GRAPE-LM-generated libraries exhibited an approximately tenfold higher retention rate than random libraries on the CD3 ϵ target and an approximately 100-fold higher retention rate than random libraries on the RBD target (Fig. 2b). Sequences generated by GRAPE-LM with the representing RNA-FM and Evo showed similar overall retention rates. Read counts in CRISmers NGS data quantitatively represent relative RNA binding activity to its target. Therefore, we further quantified sequence enrichment with both the proportion of sequences with top read counts and their average read counts to gain insights into functional contexts. The sequences from GRAPE-LM with RNA-FM showed higher enrichment in top reads and read counts than those from the model with Evo (Fig. 2c). These results align with the different pretraining data of the two LMs: RNA-FM pretrained specifically with binding-related noncoding RNAs resulted in greater possibility of producing highly enriched aptamer candidates compared to Evo pretrained with genomic data.

To directly compare computational and experimental approaches, we performed a second-round iterative CRISmers screening using results from the initial CRISmers screen. We evaluated to what extent GRAPE-LM could recapitulate manual screening efforts by analyzing the overlap between GRAPE-LM-generated candidates and the top 1,000 sequences of the highest read counts from the second-round iterative screen. Sequence similarity was quantified using Cd-hit (3-base error tolerance). We evaluated the libraries generated by GRAPE-LM sampling from the relatively high-activity and low-activity regions of the latent space under different relative numbers (normalized by diversity) of generated samples. As a control, we also evaluated the sequence library generated by AptaTrans, the best competing method in terms of generated diversity (Fig. 1i). Due to the high time cost of generating sequences using AptaTrans (approximately 18 hours for 5,000 sequences, as shown in Fig. 1j), we only compared it up to a scale of 100,000. By leveraging activity guidance from high-activity regions, GRAPE-LM (h) captures substantially more of the top 1,000 validated leads compared to AptaTrans and GRAPE-LM (l) (Fig. 2d). GRAPE-LM (l) is informed by low-activity aptamers. This enrichment efficiency enables high-activity trained GRAPE-LM to identify promising candidates using fewer samples while achieving equivalent discovery of high-potential molecules.

Benchmarking of GRAPE-LM-derived CD3 ϵ aptamer leads

To further validate the above findings, we compared GRAPE-LM-derived (post-AI) aptamer candidates against published CD3 ϵ -targeting aptamers^{34–36} (Supplementary Table 3) and those identified through pre-GRAPE-LM one-round CRISmers screening (pre-AI). Using a CRISPR–Cas-based luciferase reporter assay (Fig. 3a), we assessed the top 50 post-AI aptamers (determined by read count; Fig. 3a). Among these, 23 aptamers exhibited stronger luciferase activation, a proxy for binding activity, than the best pre-AI candidates (pre-AI Lead 1 and Lead 2; Supplementary Table 4). By contrast, manual second-round

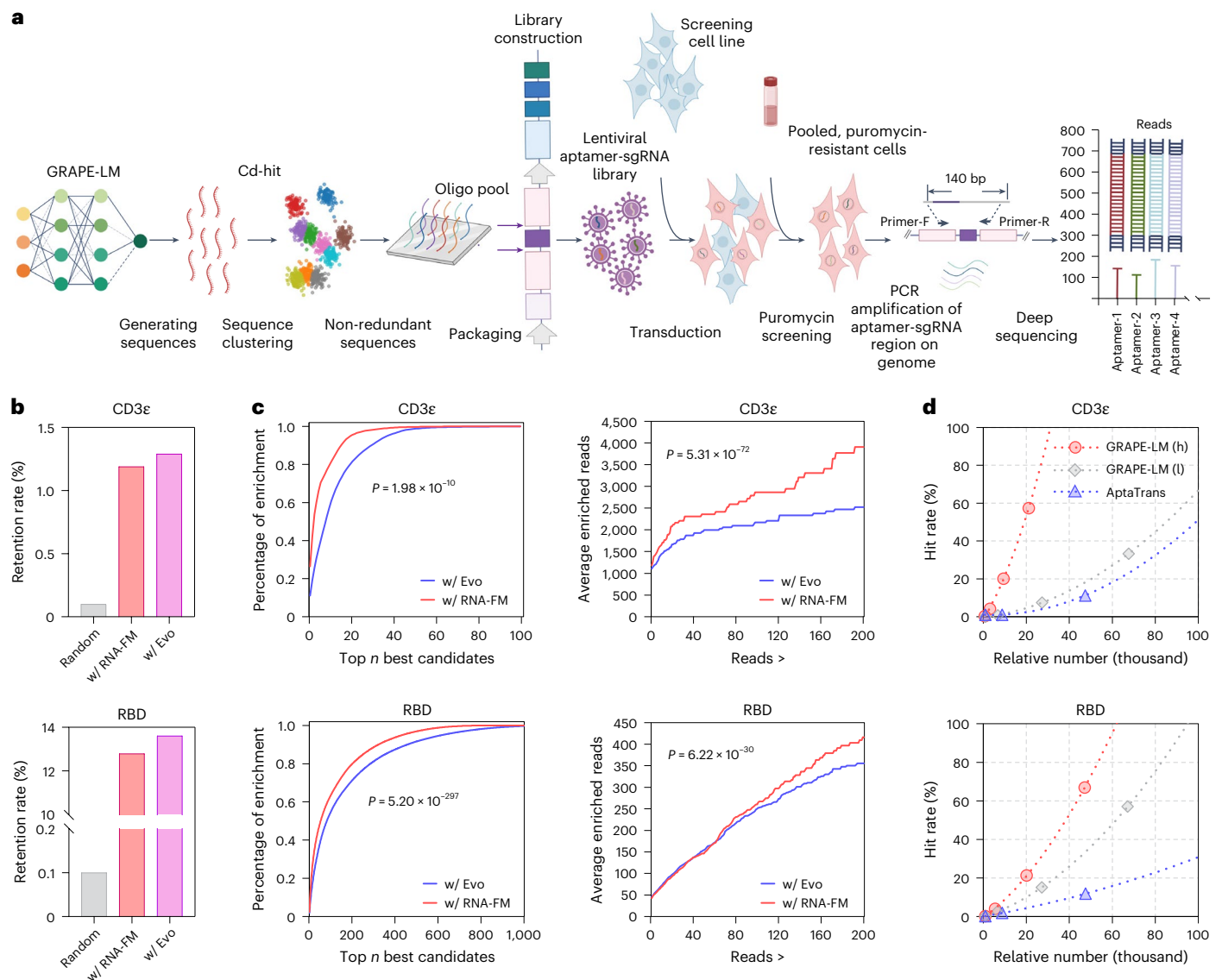


Fig. 2 | Experimental validation of RNA aptamer evolution with GRAPE-LM.

a, Schematic of the workflow of CRISpers-based assessment for sequence output libraries generated from GRAPE-LM (with RNA-FM²⁶ and Evo³¹) on two targets, respectively. To mitigate redundancy and experimental bias, we applied Cd-hit³³ with an 80% similarity threshold to establish non-redundant sequence libraries from cost-efficient evaluations of 10,000 RNA-FM and 10,000 Evo generated sequences (Supplementary Table 2). Icons in figures were created with BioRender.com. **b**, Retention rates of GRAPE-LM-generated and random sequence libraries after one round of CRISpers examination. Retention rate is defined as the fraction of enriched unique sequences relative to the total library size. **c**, Comparative analyses of CRISpers output NGS data derived from models

with varying LMs. Statistical significance was assessed using two-sided Student's *t*-tests. **d**, The hit rates that cover the top 1,000 sequences obtained in the second-round recycled CRISpers screening, by generating different numbers of aptamer candidates using GRAPE-LM. The hit rates were calculated based on sequence similarity with 3-base error tolerance using Cd-hit. The libraries generated from AptaTrans, informed by the corresponding target protein, were used as a control. The quantity of the generated sequences was normalized by diversity, with redundant, highly similar sequences excluded from the count. sgRNA, single guide RNA; GRAPE-LM (h), sampling from a relatively high-activity region; GRAPE-LM (l), sampling from a relatively low-activity region.

CRISpers screening did not produce sequences with higher affinity compared to the first-round hits (Supplementary Fig. 2). Many post-AI aptamers surpassed previously reported sequences^{35,36} derived from 10 or 16 rounds of human selection and evolution (Fig. 3a).

To further test binding affinity and specificity, we employed an independent flow cytometry assay using CD3-expressing Jurkat T cells (Supplementary Fig. 3a), an established method for evaluating aptamers targeting membrane proteins in their native conformation^{34,37,38}. Among these post-AI aptamer leads confirmed by luciferase reporter assay, we focused on the top three (post-AI-19, post-AI-34 and post-AI-38; Supplementary Table 4) in luciferase reporter assay (Fig. 3a). All three leads showed higher binding activity to Jurkat T cells

than pre-AI Lead 1 and Lead 2 (Supplementary Table 4), which represent the best results prior to GRAPE-LM involvement (Fig. 3b and Supplementary Fig. 3b).

To rigorously assess the binding affinity and specificity of the selected aptamers, we generated two independent CD3ε-knockout Jurkat monoclonal cell lines via CRISPR-Cas9 (Supplementary Fig. 4a). Successful knockout was validated through Sanger sequencing (Supplementary Fig. 4b,c) and monoclonal subcloning with genotyping (Supplementary Fig. 4d,f). Both clones were biallelic frameshift mutants: clone 1 (Jurkat CD3ε Ko 1) carried 4-base pair (bp) and 8-bp insertions, and clone 2 (Jurkat CD3ε Ko 2) carried 7-bp and 2-bp insertions. The complete loss of CD3ε expression in each monoclonal

line was further confirmed by flow cytometry (Supplementary Fig. 4e,g), establishing them as robust negative controls for subsequent specificity evaluation.

Dose-dependent binding assays conducted by flow cytometry demonstrated that the post-AI aptamers exhibit high affinity and specificity. The top candidate, CD3ε post-AI-34, showed a dissociation constant (K_d) of 53 ± 12 nM, outperforming pre-AI leads (K_d range: 360–692 nM), as well as previously reported CD3ε aptamers derived from multi-round SELEX, such as JZPO-10 (114 ± 23.9 nM) and ZUCH-1 (69 ± 15.2 nM)^{35,36} (Fig. 3c and Extended Data Fig. 4a,b). Specificity was confirmed by negligible binding signals on CD3ε-knockout cells, which were similar to background levels in negative controls. AI-designed aptamers not only display robust affinity but also achieve markedly higher specificity among three leads, notably outperforming the low specificity observed in SELEX-generated lead ZUCH-1 despite its high affinity observed in wild-type Jurkat cells. Physiologically relevant, cell-based assays using CD3ε-expressing and knockout cells enabled a head-to-head comparison of binding affinities between algorithmically generated and experimentally derived aptamers.

To evaluate the structural and sequence novelty of the three GRAPE-LM-generated aptamer candidates, we compared them against previously published nucleic acid aptamers^{35,36} (Fig. 3d using AlphaFold 3 (ref. 39) and Supplementary Fig. 5a using Vfold⁴⁰, a top model for RNA monomer prediction in CASP16 (ref. 41)). Both sequence and structure alignments revealed minimal overlap between the two groups. Then, we expanded this analysis to include the top 50 post-AI GRAPE-LM-generated candidates and the top 50 pre-AI CRISmers-derived aptamers (Fig. 3e). Both the Vfold web-server documentation and our empirical validation demonstrate that certain input samples cannot be processed due to constraints within the existing template database⁴⁰. Therefore, the structures of all these aptamers were predicted using AlphaFold 3 (ref. 39). Structural and sequence alignments again demonstrated negligible overlap between the two groups. This divergence confirms the ability of GRAPE-LM to explore novel sequence and structural spaces for aptamer evolution.

Notably, some post-AI aptamer leads (especially post-AI-15, post-AI-17, post-AI-19 and post-AI-48) exhibited structural convergence with pre-AI aptamer leads despite lacking sequence homology (Fig. 3e), implying that the model may infer conserved structural motifs critical for functionality. To further probe this, we extended sequence and structure alignment to all enriched post-AI candidates and the pre-AI top 1,000. Then, we clustered the structural alignment results based on template modeling scores (TM-scores) and arranged the corresponding sequence alignments in the same order (Supplementary Fig. 5b). Structural clustering revealed distinct high TM-score clusters, whereas the corresponding sequence alignments showed no obvious aggregation. To quantify this difference, we binarized the structural alignment results using the average TM-score as a threshold and applied the

same approach to sequence alignments using the average sequence alignment score. We then calculated Shannon entropy⁴² for the overall structural and sequence alignments as well as for the high TM-score clusters (50×50) and their corresponding sequence alignment regions. The results showed a marked reduction in entropy within the high TM-score structural clusters, whereas the entropy of the corresponding sequence regions remained nearly unchanged compared to the overall values. Taken together, these results suggest that structural convergence indeed occurs between GRAPE-LM-generated sequences and pre-AI sequences, independent of sequence similarity.

As for case studies, we visualized the sequence patterns of four post-AI top hits and their structurally matched counterparts, respectively. The results show that there are potential conserved regions within each cluster (Fig. 3g). Interestingly, post-AI-19, one of the top three post-AI leads, displayed structural homology to five pre-AI aptamer leads but without sequence similarity (Fig. 3f, Supplementary Fig. 6 and Supplementary Table 4). Molecular docking (HDOCK⁴³) of post-AI-19 with human CD3ε (Protein Data Bank (PDB): 1XIW) localized the conserved pattern to binding regions of the aptamer (Fig. 3g,h).

Validation of GRAPE-LM on RBD

We also applied the framework to evolve aptamers targeting a protein of viral origin: RBD. Sequence and structural comparisons between the top 50 post-AI GRAPE-LM-generated aptamers and the top 50 pre-AI CRISmers-derived candidates revealed minimal overlap (Fig. 4a), mirroring trends observed in the CD3ε-targeting analysis.

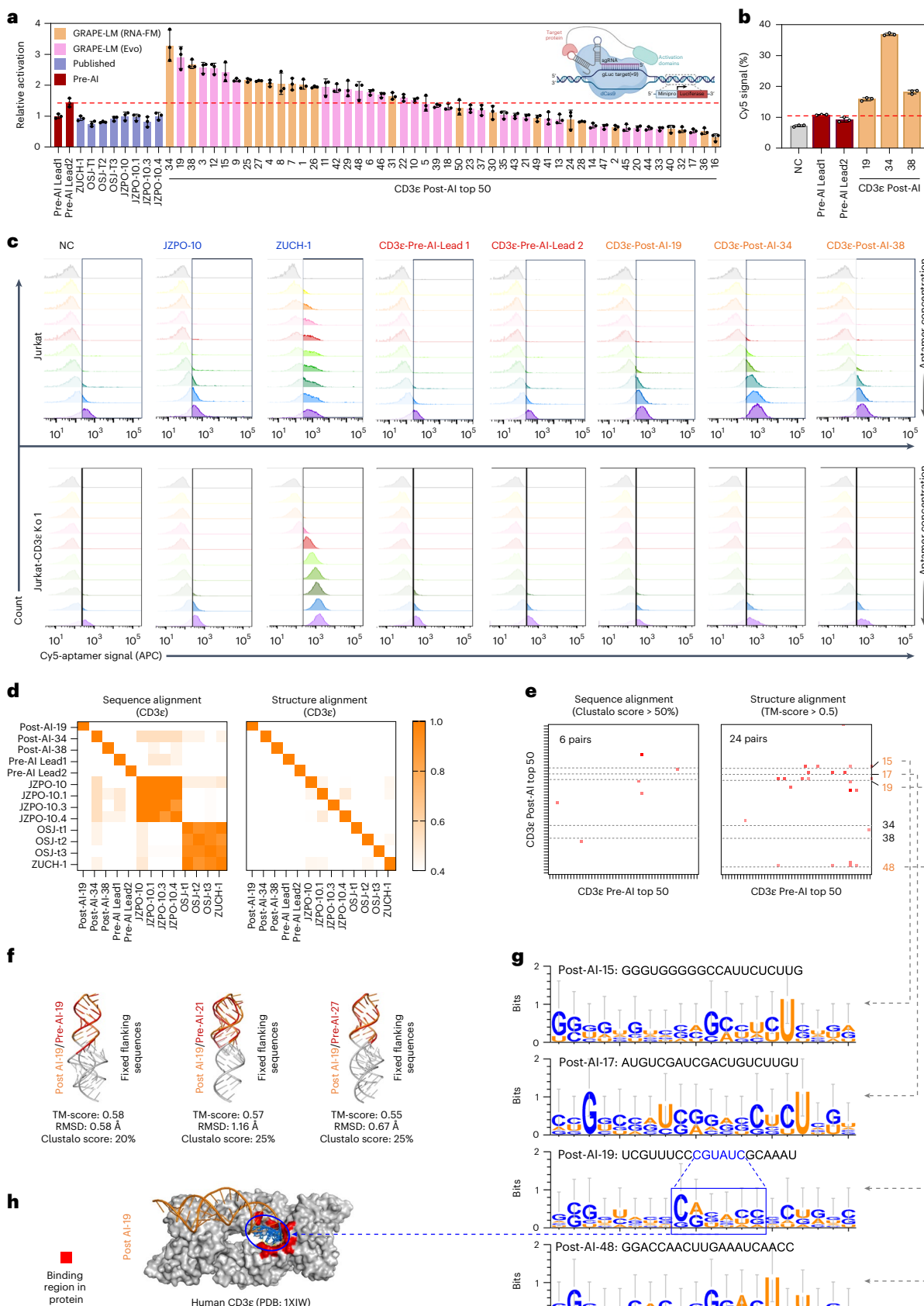
Building on the success of GRAPE-LM in generating highly active CD3ε-targeting aptamers, we functionally validated the top 20 SARS-CoV-2 RBD candidates (ranked by sequencing read abundance) using the CRISPR–Cas-based luciferase reporter assay (Fig. 4b; see Fig. 3a for illustration). More than half of the top 20 GRAPE-LM-derived aptamers outperformed two representative RNA aptamers (RBD-PB6 (ref. 44) and CoV2-6C3 (ref. 45); Supplementary Table 3), which were respectively derived from traditional 8-round and 13-round SELEX campaigns. These two benchmark aptamers were selected due to their highest level of signal in a previously analyzed set of published aptamers targeting RBD⁴¹. In addition, luciferase activation of the top 50 most frequent sequences obtained from one round of CRISmers screening prior to GRAPE-LM involvement was markedly lower than the benchmarking RBD post-AI-14 (Fig. 4b and Extended Data Fig. 5a). The effects of GRAPE-LM were visually characterized through parallel green fluorescent protein (GFP) reporter assays conducted by substituting luciferase with GFP (Fig. 3a), enabling quantitative affinity evaluation via flow cytometry. When benchmarking the pre-GRAPE-LM control library against its AI-engineered counterpart, GRAPE-LM, flow cytometry analyses revealed substantially higher fluorescence intensity in the GRAPE-LM output population, indicative of successful enrichment for high-expressing cells (Extended Data Fig. 5b).

Fig. 3 | Benchmarking GRAPE-LM-derived aptamer leads targeting CD3ε. a, Schematic illustration of the mechanism of action of the CRISPR–Cas-based luciferase activation system to quantify aptamer binding activity and the benchmarking results of previously published pre-AI and GRAPE-LM-generated leads. Data plot mean with s.d. ($n = 3$ biological replicates). Icons in figures were created with BioRender.com. **b,** Flow cytometry results targeting CD3ε-positive human T-cell-derived Jurkat cells. Data show mean with s.d. ($n = 3$ biological replicates). **c,** Dose-dependent flow cytometry analyses of the Cy5-labeled CD3ε aptamers on Jurkat cells. CD3ε-knockout cells (Jurkat CD3ε Ko 1 (this panel) and CD3ε Ko 2 (Extended Data Fig. 4a)) serve as negative control to assess binding specificity. Data show representative flow cytometry plots from three independent experiments ($n = 3$ biological replicates). **d,** The comparison results of the sequence and structural similarity between the aptamer candidates confirmed by flow cytometry and existing nucleic acid aptamers. The structural similarity was calculated using US-align⁴⁶, and sequence similarity was calculated

using Clustal Omega⁴². The fixed flanking sequences of aptamers were not used in sequence or structure alignment. **e,** The results of the sequence and structure alignment between top 50 GRAPE-LM-generated (post-AI) candidates and top 50 pre-AI CRISmers-screened aptamers (ranked by sequencing read abundance). All the structures were predicted using AlphaFold 3 (ref. 39). **f,** Illustrations of the three-dimensional structures of CD3ε post-AI-19 and its top three structurally matched counterparts (Supplementary Table 4) using PyMOL. Each pair shares structural similarity but no sequence homology. Neither sequence nor structure alignment scores took account of the fixed flanking sequences. **g,** Visualization of sequence patterns using WebLogo 3 (ref. 63). Each logo visualizes the multiple sequences derived from a post-AI lead and its structurally matched counterparts. **h,** Visualization of the binding conformation and binding region of post-AI-19 with CD3ε target (PDB: 1XIW) predicted by HDOCK⁴³. NC, negative control RNA aptamer; RMSD, root mean square deviation; sgRNA, single guide RNA; Clustalo, Clustal Omega.

We next employed microscale thermophoresis (MST; Extended Data Fig. 6a)⁴⁶ to quantitatively evaluate the binding affinity of aptamer leads generated by GRAPE-LM against both pre-AI sequences and independently validated benchmarks from previous publications

(Supplementary Table 3)^{11,44,45,47–52}. Three GRAPE-LM-derived aptamers—post-AI-4, post-AI-12 and post-AI-14—which had exhibited high relative activation in luciferase reporter assays and shared sequence or structural similarity with pre-AI leads (pre-AI-10 and pre-AI-37) (Fig. 4c,d



and Supplementary Table 4), demonstrated superior binding affinity compared to their pre-AI counterparts ($K_d = 61.7 \pm 7.4$ nM (post-AI-14) versus 157.7 ± 37.6 nM (pre-AI-10); $K_d = 37.3 \pm 5.4$ nM (post-AI-12) and 66.8 ± 10.4 nM (post-AI-4) versus 85.6 ± 26.6 nM (pre-AI-37)). Good affinity was also observed in a post-AI lead (post-AI-2) ($K_d = 85 \pm 13.4$ nM) with low sequence or structural similarity to any pre-AI aptamers, suggesting that GRAPE-LM may be able to explore novel sequence space beyond prior library constraints. Moreover, among the previously reported SELEX benchmarks, K_d values range from 64 ± 4.3 nM to 1.96 ± 0.08 μ M (Extended Data Fig. 6b), with DSA1N1 recording the highest affinity ($K_d = 64 \pm 4.3$ nM). In comparison, our top GRAPE-LM-generated RBD aptamers, such as post-AI-12 ($K_d = 37.3 \pm 5.4$ nM), achieved affinities that were similar to or superior than the best among these SELEX benchmarks (Extended Data Fig. 6b)^{11,44,45,47–52}. The top lead post-AI-12 exhibited higher binding affinity than both the second-round ($K_d = 74.1 \pm 16$ nM) and the fifth-round ($K_d = 55 \pm 13.6$ nM) CRISmers-derived leads (Extended Data Fig. 6c), highlighting the potential of AI-driven design to replace multi-round experimental iterations, including our CRISPR-based intracellular screening.

To further confirm the novelty of the three GRAPE-LM-generated aptamer candidates, we also compared sequence and structural features between them and the RBD aptamer leads previously published (Supplementary Table 3)^{11,44,45,47–52}. The results revealed distinct differences in both sequence composition and three-dimensional structural conformation between the GRAPE-LM-generated candidates and previously reported nucleic acid aptamers in SELEX studies, thus confirming the structural and sequence novelty of our best leads compared to previous efforts (Fig. 4e using AlphaFold 3 and Supplementary Fig. 7a using Vfold).

In this study, we observed structural convergence between GRAPE-LM-generated sequences and pre-AI designs in both RBD (Supplementary Fig. 7b) and CD3 ϵ (Supplementary Fig. 5b), suggesting that conserved motifs drive functional binding. To identify functional motifs within our lead aptamers targeting CD3 ϵ and RBD, we used RNAFold (ViennaRNA)⁵³ to predict their secondary structures (Supplementary Fig. 8). All these functional aptamers contain structural motifs such as internal loops that might serve as critical determinants for their binding aptamer function. As for case studies, binding residues within these loops in CD3 ϵ post-AI-38 and RBD post-AI-14, one for each target, were mapped computationally using an integrated approach combining AlphaFold 3 and HDOCK (Extended Data Fig. 7). Functional validation was performed via site-directed mutagenesis, introducing transversions between purines and pyrimidines (A $\leftrightarrow$ C and G $\leftrightarrow$ U) at key residues to maximize structural disruption through altered base pairing and local architectural distortion⁵⁴. Mutagenesis of the predicted binding sites within the internal loops severely impaired binding affinity for both aptamers (Extended Data Fig. 7). In CD3 ϵ post-AI-38, changing any one of these six bases impaired binding, increasing the K_d (Extended Data Fig. 7a). The substitution of the final cytosine for an adenine (C $\rightarrow$ A) had the strongest effect on a single mutant, markedly reducing affinity ($K_d = 1.7 \pm 0.6$ μ M). Disrupting all six nucleotides simultaneously resulted in the greatest impairment,

with the K_d rising to 3.1 ± 1.1 μ M. Notably, substituting merely a single predicted contact site in RBD post-AI-14's internal loop region caused profound loss of target binding capability (Extended Data Fig. 7b). These data underscore the essential role of these specific structural features in mediating aptamer–target interactions.

We next evaluated the efficacy of GRAPE-LM-evolved aptamer candidates in neutralizing SARS-CoV-2 pseudovirus (Fig. 4f). Focusing on post-AI-12 and post-AI-14—the two top-ranked leads identified by MST binding affinity (Fig. 4d)—we observed robust, dose-dependent inhibition of pseudoviral infection compared to an irrelevant control sequence. These AI-generated aptamers exhibited broad-spectrum neutralization across multiple SARS-CoV-2 variants of concern, including wide-type, Delta and Omicron, consistent with previous studies demonstrating pan-inhibition capabilities of previous aptamer leads^{11,55}.

GRAPE-LM can target an intracellular disordered protein, c-Myc
Having demonstrated proof-of-concept efficacy on CD3 ϵ and RBD targets, we extended GRAPE-LM testing to c-Myc⁵⁶—an intracellular disordered oncogenic transcription factor—to assess its broader applicability across diverse molecular targets. This rigorous test case probes platform robustness for challenging intracellular targets. The results on c-Myc were consistent with the findings obtained for CD3 ϵ and RBD in terms of computational performance metrics (Fig. 5a–d). The CAE framework outperformed VAE, masked language modeling and diffusion-based generative approaches (Fig. 5a). Incorporating RNA LMs enhanced sequence generation quality (Fig. 5b). Furthermore, the overall performance of GRAPE-LM measured by multi-metrics including computational efficiency (Fig. 1j) consistently outperformed four existing nucleic acid aptamer generation methods as benchmarks (Fig. 5c). The effectiveness of the activity guidance mechanism in GRAPE-LM was confirmed while targeting c-Myc (Fig. 5d). This validation was achieved by analyzing precision-recall curves, comparing it against a model without guidance and calculating the Pearson's correlation coefficient between the predicted pseudo-activity scores and the ground truth data derived from CRISmers screening.

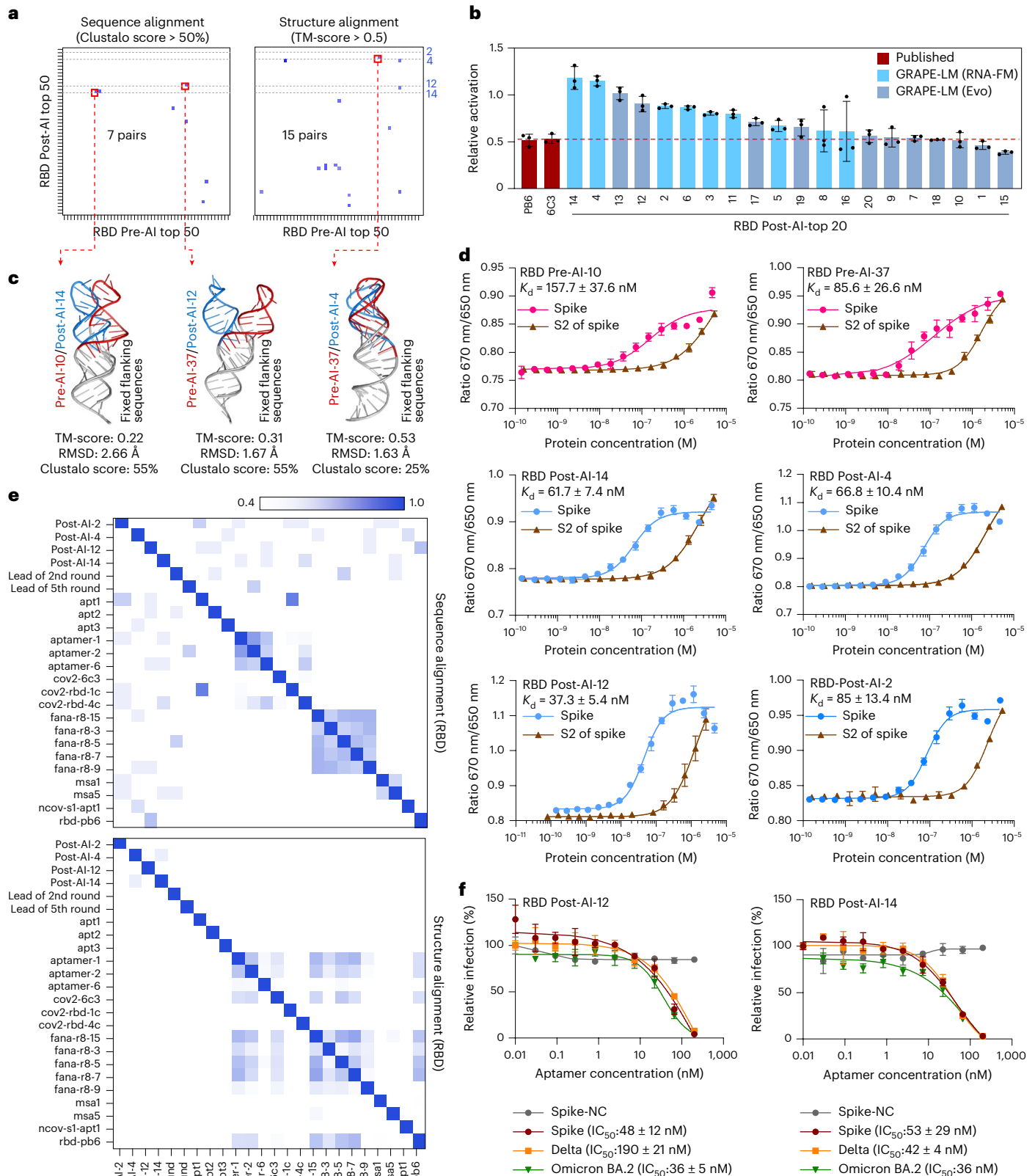
We further conducted wet lab experiments to validate the role of activity guidance to GRAPE-LM and evaluate the performance of different generative frameworks and methods (Fig. 5e–g), using the CRISmers-based pooled assay (Fig. 2a). The results revealed that sequences sampled from the relatively high-activity regions of GRAPE-LM exhibited a retention rate twice as high as those sampled from low-activity regions of the GRAPE-LM latent space (Fig. 5e), consistent with their predicted pseudo scores (Supplementary Fig. 9a) and latent spatial distribution (Supplementary Fig. 9b). This advantage remained significant across different readout thresholds, measured by not only retention rate but also average read enrichment (Fig. 5g). When we applied an additional filtration step using high pseudo-activity scores predicted by GRAPE-LM (top 1,000 of 100,000 GRAPE-LM-generated sequences, ranked by the pseudo-activity scores predicted from GRAPE-LM), the retention rate of the selected sequences on the c-Myc target increased by nearly 17 folds (Fig. 5f). These findings further validate the functional role of activity guidance in GRAPE-LM.

Fig. 4 | Validation of GRAPE-LM-derived aptamer leads targeting RBD. **a**, The results of the sequence and structure alignment between top 50 GRAPE-LM-generated (post-AI) candidates and top 50 pre-AI CRISmers aptamer leads (ranked by sequencing read abundance). All the structures were predicted using AlphaFold 3 (ref. 39). The structural similarity was calculated using US-align⁶¹, and sequence similarity was calculated using Clustal Omega⁶². The fixed flanking sequences were not used in sequence or structure alignment. **b**, Results of the luciferase reporter assay (Fig. 3a) for top 20 post-AI aptamer candidates (ranked by sequencing read abundance). Data plot mean with s.d. ($n = 3$ biological replicates). **c**, Visualization of the three-dimensional structures of the post-AI aptamer candidates (Supplementary Table 4) that share sequence or structural similarity with pre-AI leads (Supplementary Table 4) and elicit high level of activation in the luciferase

reporter assay. Neither sequence nor structure alignment scores took account of the fixed flanking sequences. **d**, Results of K_d value determination for the above two pre-AI aptamers and four post-AI aptamer leads (**b**) based on MST (Extended Data Fig. 6a). The RBD is located within the S1 subunit of the SARS-CoV-2 spike protein. The S2 subunit of spike was employed as a negative control in MST to assess the specificity of the aptamer. Data show mean $\pm$ s.d. ($n = 3$ biological replicates). **e**, The comparative results of the sequence and structural similarity between the aptamer candidates confirmed by MST (**d**) and existing nucleic acid aptamers. **f**, Neutralizing activity of leading aptamers against the pseudo-SARS-CoV-2 variants. Data show mean $\pm$ s.d. ($n = 3$ biological replicates). IC₅₀, half-maximal inhibitory concentration; NC, negative control RNA aptamer; RMSD, root mean square deviation; Clustalo, Clustal Omega.

In comparison, although the diffusion framework can generate more novel and diverse RNA sequences (Figs. 1b and 5a), the lack of effective activity guidance makes it difficult to ensure high activity (Figs. 1c and 5a), resulting in only limited retention rates (Fig. 5e).

Similarly, no sequence generated by RaptGen was identified. Masked language modeling framework achieved higher retention than diffusion but still showed substantially lower performance compared to GRAPE-LM's CAE framework. GRAPE-LMs with multiple RNA-based LMs



have achieved higher retention rates and outperformed the other four high-throughput AI methods as well as RhoDesign, a low-throughput inverse folding-based generative model. These results are consistent with the computational results (Fig. 5a–d), underscoring the effectiveness of GRAPE-LM.

To further evaluate the success rate of GRAPE-LM, we measured the relative activation of the top 100 GRAPE-LM-derived (with various RNA LMs) c-Myc aptamer candidates via a comparative examination of binding signal using the CRISPR–Cas-based luciferase reporter assay (Fig. 3a). Notably, 30 of the top 100 GRAPE-LM-derived c-Myc aptamers (ranked by sequencing read abundance) outperformed two high-affinity SELEX-derived aptamers from previous studies (c-Myc MA9 and c-Myc MA9C1; Supplementary Table 3)⁵⁷ and two top pre-AI leads (Fig. 5h, Supplementary Fig. 10 and Supplementary Table 4). Higher affinity was confirmed for all seven GRAPE-LM-generated aptamer leads examined through MST analysis. With K_d values ranging from 264 nM to 356 nM, they outperformed the leading pre-AI candidates (386–426 nM) and also surpassed the best previously reported SELEX aptamer, MA9C1, which served as our benchmark (662 ± 78 nM) (Fig. 5i). All seven GRAPE-LM-generated aptamer leads share minimal sequence or structural similarity between the two pre-AI leads and two published aptamers targeting c-Myc (Supplementary Fig. 11), again verifying the novelty of our post-AI leads. The successful generation of high-performing c-Myc aptamer leads by multiple distinct RNA LMs, all outperforming both SELEX benchmarks and pre-AI sequences, demonstrates the compatibility of GRAPE-LM with diverse RNA LMs.

Discussion

In this study, we introduce GRAPE-LM, a generative AI framework designed to accelerate the evolution of RNA aptamers (Extended Data Fig. 8). Experimental validation across three distinct targets—human CD3ε, c-Myc and SARS-CoV-2 spike RBD—demonstrated the ability of GRAPE-LM to obtain superior binders through just one round of evolution compared to aptamer leads reported previously that were identified after recycling for 6–16 rounds (Supplementary Table 3). Specifically, our CRISmers + GRAPE-LM pipeline comprises two evolutionary phases: physical evolution within cells followed by digital optimization via LMs. This bifurcated approach fundamentally differs from traditional SELEX, which relies exclusively on real-world

selection without computational enrichment. This paves the way for highly efficient creation of functional aptamers, eliminating the need for labor-intensive and time-consuming rounds of selection by human effort.

Recent advances in protein evolution have highlighted the power of protein LMs combined with activity guidance⁵. Although this aligns with our foundational concept, our GRAPE-LM method differs from established frameworks, such as EVOLVEPro⁵. Instead of relying on iterative site-directed mutagenesis, GRAPE-LM generates novel RNA sequences de novo through latent space sampling, bypassing the stepwise mutation approaches in EVOLVEPro. Moreover, GRAPE-LM addresses critical limitations in recently published RNA aptamer design methods^{13–15,19,20} by combining a CAE with activity-guided constraints derived from experimental screening data. Unlike VAEs used in RaptGen, which introduce randomness into sequence generation and risk, producing repetitive or non-functional outputs, GRAPE-LM's activity-guided CAE ensures that generated sequences remain functionally relevant while exploring novel sequence spaces. This approach prevents the model from converging on limited sequence patterns, a common challenge in VAE-based methods^{58–60}. Although diffusion frameworks show promise for peptide generation²⁵ and excellent diversity in generated RNA aptamers¹⁴, they are presently challenged to perform activity-guided generation. Thus, the diffusion framework prioritizes sequence diversity and novelty, but its latent space constraints often limit functional relevance, resulting in a low retention rate. By contrast, GRAPE-LM's activity-guided CAE balances novelty with functional optimization, enabling the generation of sequences that are both diverse and biologically rational. This advantage is particularly evident in low-resource scenarios, where GRAPE-LM requires minimal experimental data—only one round of CRISmers screening out of random sequence library—to generate high-affinity aptamers.

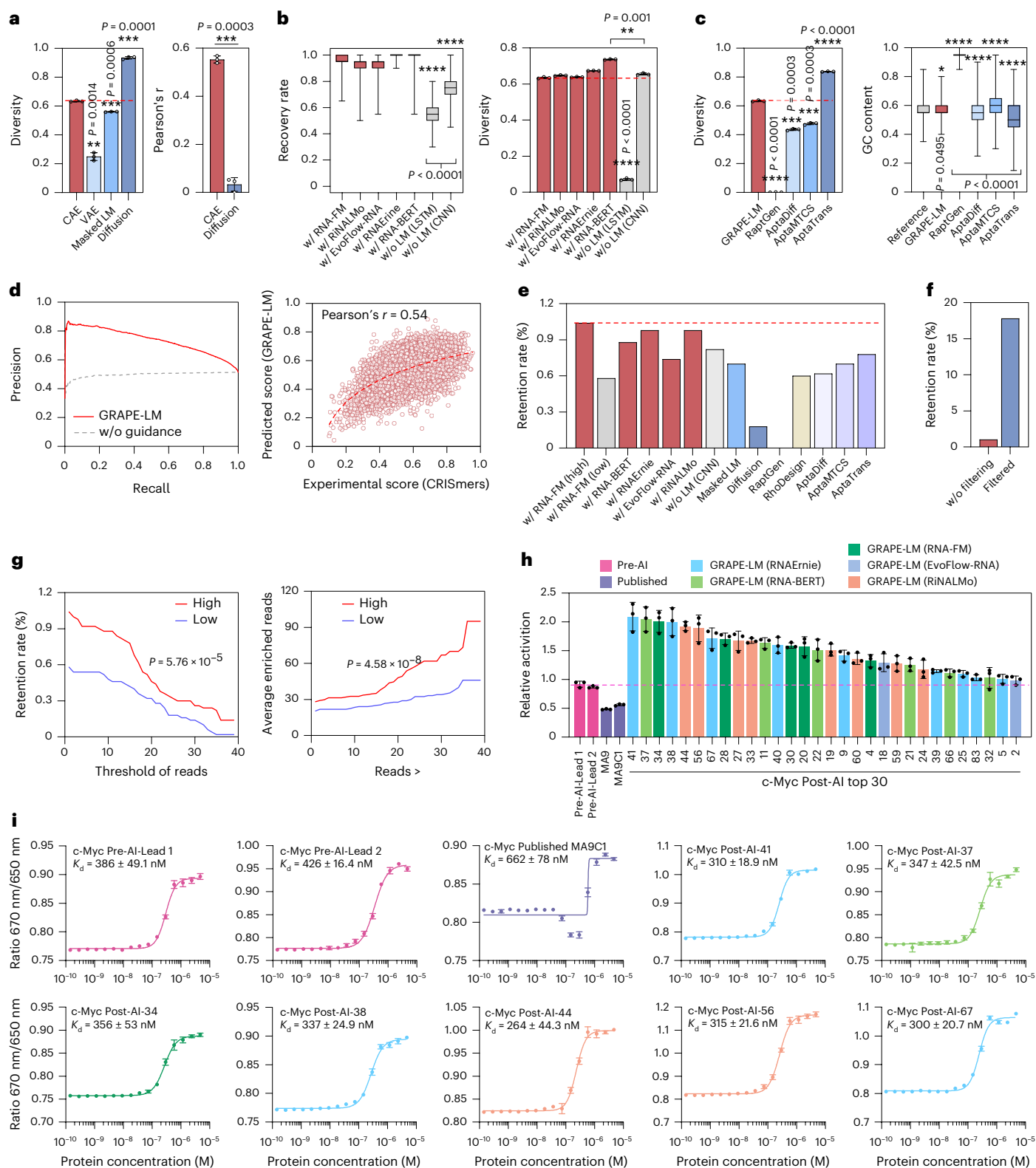
GRAPE-LM circumvents the need for structural data, a requirement for tools such as RhoDesign¹⁵, which depends on prior knowledge of RNA structures—a limitation given the challenges in accurate RNA structure determination. By contrast, no prior knowledge is needed in CRISmers + GRAPE-LM evolution, as a random sequence library is used for CRISmers screening. Approaches such as Apta-MCTS¹⁹, AptaTrans²⁰ and AptaBLE²¹, which combine aptamer–protein interaction prediction models with reinforcement learning for target protein-guided

Fig. 5 | Evaluations of GRAPE-LM on human c-Myc. **a**, Comparison of CAE with different generative frameworks for generating RNA aptamers targeting c-Myc. The generated diversity was calculated based on the samples generated by a model based on the same seed sequences obtained from one-round CRISmers targeting c-Myc. The effectiveness of activity guidance was measured by Pearson's correlation coefficient (Pearson's r) between the predicted pseudo-activity score and the ground truth derived from CRISmers on c-Myc. Data show mean with s.d. ($n = 3$ experimental replicates). Statistical analyses were conducted using a two-sided Student's t -test ($***P < 0.001$; $**P < 0.01$). **b**, The ablation study of the RNA LMs on GRAPE-LM. The recovery rate measures the ability of a model to reconstruct sequences in the test set ($n = 8,053$ samples). Diversity plot shows mean with s.d. ($n = 3$ experimental replicates). Statistical analyses were conducted using a two-sided Student's t -test ($****P < 0.0001$; $***P < 0.001$). **c**, Comparison of computational performance between GRAPE-LM and other methods. Diversity plot shows mean with s.d. ($n = 3$ experimental replicates). Statistical analyses were conducted using a two-sided Student's t -test ($***P < 0.001$; $**P < 0.01$). Generation rationality was measured by GC content ($n = 5,000$ generated samples). The top 2,000 leads from pre-AI were used as a reference. Statistical significance of GC content was assessed by unpaired Welch's two-sided t -test ($****P < 0.0001$; $*P < 0.05$). **d**, The precision-recall curve of predicted pseudo-activity score on the test sets illustrates the model's hit rate under different readout thresholds. The random control refers to a model that uses a prediction module randomly initialized without activity guidance training ($w = 0$). The scatter plot shows GRAPE-LM activity guidance effect on the c-Myc test set. The y axis represents the activity score predicted by GRAPE-LM; the x axis represents the pseudo-activity score calculated based

on CRISmers validation results (Fig. 2a). The Pearson's correlation coefficient was used as an evaluation index for the effectiveness of GRAPE-LM activity guidance. **e**, CRISmers-based wet lab comparative assessment (Fig. 2a) of GRAPE-LM for effectiveness of activity guidance (RNA-FM high versus low) and for benchmarking with alternative frameworks and existing methods. The retention rate is defined as the proportion of enriched unique sequences relative to the total number of sequences in the library. **f**, Investigation of pseudo-activity guidance on c-Myc by comparing the retention rates of filtered sequences using the pseudo-activity scores predicted by GRAPE-LM (selected top 1,000 from 100,000 generated sequences) and sequences without filtering. **g**, Analysis of the enriched sequences from latent regions of relatively high and low activity. Retention rates and average enriched reads at different readout thresholds are shown. The statistical analyses were conducted using a two-sided Student's t -test. **h**, Luciferase reporter assay results of the top 30 post-AI aptamer candidates. Aptamers ranked top 100 by sequencing read abundance in CRISmers were examined. The SELEX-derived aptamer leads MA9 and MA9C1 (ref. 57), along with pre-AI Lead 1/2, are included as benchmark controls. Aptamers are color coded according to their source LM. Data show mean with s.d. ($n = 3$ biological replicates). **i**, Results of K_d value determination (based on MST) for the above seven post-AI aptamer leads, two pre-AI leads and a published aptamer targeting c-Myc. Data are presented as mean $\pm$ s.d. ($n = 3$ biological replicates). The rectangular boxes in the box plot (**b** and **c**) represent the range of quartiles of the data, where the center line represents the median and the upper and lower edges represent the upper and lower quartiles, respectively. The line above the rectangular box represents the maximum value of the data; the line below represents the minimum value of the data.

aptamer design, represent a promising direction for future research. These methods enable de novo aptamer design without explicit activity labels and are conceptually complementary to GRAPE-LM. However, their reliability is currently constrained by the limited availability of high-quality aptamer–protein interaction data. Nonetheless, this remains a highly worthwhile and forward-looking research avenue, especially as advances in data availability and modeling technologies continue to accelerate.

The success of GRAPE-LM hinges on its integration with high-fidelity experimental screening systems such as CRISMsers. Beyond their native DNA-editing function, CRISPR–Cas tools here act as high-fidelity data generators, enabling the capture of intracellular biological processes that traditional in vitro methods such as SELEX cannot replicate. Although the library capacity of CRISMsers is inherently limited by practical constraints, such as cellular delivery efficiency and the scale of cell culture, GRAPE-LM compensates by extrapolating



from a limited starting library to explore broader sequence spaces. This synergy explains why GRAPE-LM consistently identified high-affinity aptamers for diverse targets—both human and viral, cell surface and intracellular and structured and disordered—even when starting from a single round of screening on random libraries.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-026-03007-5>.

References

- Zhou, J. & Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.* **16**, 181–202 (2017).
- Zhang, J., Lang, M., Zhou, Y. & Zhang, Y. Predicting RNA structures and functions by artificial intelligence. *Trends Genet.* **40**, 94–107 (2024).
- Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
- Zhang, H. et al. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396–403 (2023).
- Jiang, K. et al. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science* **387**, eadr6006 (2025).
- Bunka, D. H. J. & Stockley, P. G. Aptamers come of age – at last. *Nat. Rev. Microbiol.* **4**, 588–596 (2006).
- Zhang, Y., Juhas, M. & Kwok, C. K. Aptamers targeting SARS-CoV-2: a promising tool to fight against COVID-19. *Trends Biotechnol.* **41**, 528–544 (2023).
- Zhang, Y., Lai, B. S. & Juhas, M. Recent advances in aptamer discovery and applications. *Molecules* **24**, 941 (2019).
- Vargas-Montes, M. et al. Enzyme-linked aptamer assay (ELAA) for detection of toxoplasma ROP18 protein in human serum. *Front. Cell. Infect. Microbiol.* **9**, 386 (2019).
- Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
- Zhang, J. et al. Repurposing CRISPR/Cas to discover SARS-CoV-2 detecting and neutralizing aptamers. *Adv. Sci.* **10**, 2300656 (2023).
- Su-Tobon, Q. et al. CRISPR-Hybrid: a CRISPR-mediated intracellular directed evolution platform for RNA aptamers. *Nat. Commun.* **16**, 595 (2025).
- Iwano, N., Adachi, T., Aoki, K., Nakamura, Y. & Hamada, M. Generative aptamer discovery using RaptGen. *Nat. Comput. Sci.* **2**, 378–386 (2022).
- Wang, Z. et al. Aptadiff: de novo design and optimization of aptamers based on diffusion models. *Brief. Bioinform.* **25**, bbae517 (2024).
- Wong, F. et al. Deep generative design of RNA aptamers using structural predictions. *Nat. Comput. Sci.* **4**, 829–839 (2024).
- Zhou, X. et al. ProRefiner: an entropy-based refining strategy for inverse protein folding with global graph attention. *Nat. Commun.* **14**, 7434 (2023).
- Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
- Ren, M., Yu, C., Bu, D. & Zhang, H. Accurate and robust protein sequence design with CarbonDesign. *Nat. Mach. Intell.* **6**, 536–547 (2024).
- Lee, G., Jang, G. H., Kang, H. Y. & Song, G. Predicting aptamer sequences that interact with target proteins using an aptamer-protein interaction classifier and a Monte Carlo tree search approach. *PLoS ONE* **16**, e0253760 (2021).
- Shin, I. et al. Aptatrans: a deep neural network for predicting aptamer-protein interaction using pretrained encoders. *BMC Bioinformatics* **24**, 447 (2023).
- Patel, S. et al. AptABLE: an enhanced deep learning platform for aptamer protein interaction prediction and design. In *Machine Learning for Structural Biology Workshop, NeurIPS 2024* https://www.mlsb.io/papers_2024/AptABLE:_An_Enhanced_Deep_Learning_Platform_for_Aptamer_Protein_Interaction_Prediction_and_Design.pdf (2024).
- Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
- Nithin, C., Kmiecik, S., Błaszczuk, R., Nowicka, J. & Tuszyńska, I. Comparative analysis of RNA 3D structure prediction methods: towards enhanced modeling of RNA–ligand interactions. *Nucleic Acids Res.* **52**, 7465–7486 (2024).
- Chen, L. T. et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-025-02761-2> (2025).
- Torres, M. D. T., Chen, L. T., Wan, F., Chatterjee, P. & de la Fuente-Nunez, C. Generative latent diffusion language modeling yields anti-infective synthetic peptides. *Cell Biomater.* <https://doi.org/10.1016/j.celbio.2025.100183> (2025).
- Shen, T. et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat. Methods* **21**, 2287–2298 (2024).
- Akiyama, M. & Sakakibara, Y. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genom. Bioinform.* **4**, lqac012 (2022).
- Wang, N. et al. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nat. Mach. Intell.* **6**, 548–557 (2024).
- Patel, S., Peng, F. Z., Fraser, K., Chatterjee, P. & Yao, S. EvoFlow-RNA: generating and representing non-coding RNA with a language model. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.02.25.639942> (2025).
- Penić, R. J., Vlašić, T., Huber, R. G., Wan, Y. & Šikić, M. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. *Nat. Commun.* **16**, 5671 (2025).
- Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
- Ishida, R. et al. RaptRanker: in silico RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information. *Nucleic Acids Res.* **48**, e82 (2020).
- Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
- Freage, L., Jamal, D., Williams, N. B. & Mallikaratchy, P. R. A homodimeric aptamer variant generated from ligand-guided selection activates the T cell receptor cluster of differentiation 3 complex. *Mol. Ther. Nucleic Acids* **22**, 167–178 (2020).
- Zumrut, H. et al. Ligand-guided selection with artificially expanded genetic information systems against TCR-CD3ε. *Biochemistry* **59**, 552–562 (2020).
- Zumrut, H. E. et al. Integrating ligand-receptor interactions and in vitro evolution for streamlined discovery of artificial nucleic acid ligands. *Mol. Ther. Nucleic Acids* **17**, 150–163 (2019).
- Raddatz, M.-S. L. et al. Enrichment of cell-targeting and population-specific aptamers by fluorescence-activated cell sorting. *Angew. Chem. Int. Ed. Engl.* **47**, 5190–5193 (2008).
- Nakhjavani, M. et al. A flow cytometry-based cell surface protein binding assay for assessing selectivity and specificity of an anticancer aptamer. *J. Vis. Exp.* **187**, e64304 (2022).
- Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

40. Li, J., Zhang, S., Zhang, D. & Chen, S.-J. Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics* **38**, 4042–4043 (2022).
41. Kretsch, R. C. et al. Functional relevance of CASP16 nucleic acid predictions as evaluated by structure providers. *Proteins* **94**, 51–78 (2026).
42. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
43. Yan, Y., Tao, H., He, J. & Huang, S.-Y. The HDock server for integrated protein–protein docking. *Nat. Protocols* **15**, 1829–1852 (2020).
44. Valero, J. et al. A serum-stable RNA aptamer specific for SARS-CoV-2 neutralizes viral entry. *Proc. Natl Acad. Sci. USA* **118**, e2112942118 (2021).
45. Sun, M. et al. Aptamer blocking strategy inhibits SARS-CoV-2 virus infection. *Angew. Chem. Int. Ed. Engl.* **60**, 10266–10272 (2021).
46. Bartoschik, T. et al. Near-native, site-specific and purification-free protein labeling for quantitative protein interaction analysis by MicroScale Thermophoresis. *Sci. Rep.* **8**, 4977 (2018).
47. Song, Y. et al. Discovery of aptamers targeting the receptor-binding domain of the SARS-CoV-2 spike glycoprotein. *Anal. Chem.* **92**, 9895–9900 (2020).
48. Li, J. et al. Diverse high-affinity DNA aptamers for wild-type and B.1.1.7 SARS-CoV-2 spike proteins from a pre-structured DNA library. *Nucleic Acids Res.* **49**, 7267–7279 (2021).
49. Liu, X. et al. Neutralizing aptamers block S/RBD–ACE2 interactions and prevent host cell infection. *Angew. Chem. Int. Ed. Engl.* **60**, 10273–10278 (2021).
50. Yang, G. et al. Identification of SARS-CoV-2-against aptamer with high neutralization activity by blocking the RBD domain of spike protein 1. *Signal Transduct. Target Ther.* **6**, 227 (2021).
51. Alves Ferreira-Bravo, I. & DeStefano, J. J. Xeno-nucleic acid (XNA) 2'-fluoro-arabino nucleic acid (FANA) aptamers to the receptor-binding domain of SARS-CoV-2 S protein block ACE2 binding. *Viruses* **13**, 1983 (2021).
52. Saify Nabiabad, H., Amini, M. & Demirdas, S. Specific delivering of RNAi using spike's aptamer-functionalized lipid nanoparticles for targeting SARS-CoV-2: a strong anti-Covid drug in a clinical case study. *Chem. Biol. Drug Des.* **99**, 233–246 (2022).
53. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
54. Guo, C. et al. Transversions have larger regulatory effects than transitions. *BMC Genomics* **18**, 394 (2017).
55. Sun, M. et al. Spherical neutralizing aptamer inhibits SARS-CoV-2 infection and suppresses mutational escape. *J. Am. Chem. Soc.* **143**, 21541–21548 (2021).
56. Dang, C. V., Reddy, E. P., Shokat, K. M. & Soucek, L. Drugging the 'undruggable' cancer targets. *Nat. Rev. Cancer* **17**, 502–508 (2017).
57. Wang, Y. et al. Antitumor effect of anti-c-Myc aptamer-based PROTAC for degradation of the c-Myc protein. *Adv. Sci.* **11**, 2309639 (2024).
58. He, J., Spokoyny, D., Neubig, G. & Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. In *The Seventh International Conference on Learning Representations* <https://openreview.net/pdf/47f79f4015dbabc7f2eab6e432cddf975cf1c486.pdf> (ICLR, 2019).
59. Dieng, A. B., Kim, Y., Rush, A. M. & Blei, D. M. Avoiding latent variable collapse with generative skip models. In *Proc. Twenty-Second International Conference on Artificial Intelligence and Statistics* (eds Kamalika, C. & Masashi, S.) 2397–2405 (PMLR, 2019).
60. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
61. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115 (2022).
62. Madeira, F. et al. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res.* **52**, W521–W525 (2024).
63. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2026

Methods

Datasets

The data used in this study were produced by the CRISPR–Cas-based aptamer screening system CRISmers, as proposed in our previous study¹¹. Unlike other SELEX screening systems, CRISmers is an intracellular RNA aptamer screening system that selects aptamers within the cellular environment, specifically binding to a target protein. For a given target, we used the output data from the first round of CRISmers screening. For an objective evaluation, we first removed the sequences that shared similarity with previously published aptamers of a target from the unique sequences. The enrichment frequency of unique sequences is often used as an indicator to measure the activity levels of candidate aptamers. Therefore, this study designed a pseudo-activity score as activity guidance information by transforming the enrichment frequency of unique sequences to a decimal between 0 and 1. To reduce redundancy and the impact of these noises introduced by polymerase chain reaction (PCR) amplification artifacts, procedural inconsistencies and variations in transfection efficiency, we used the Cd-hit program to perform sequence similarity clustering on unique sequences using a threshold of 0.8. We then aggregated read counts across multiple sequences within a cluster to compute the pseudo-activity score (Extended Data Fig. 1b). Only sequences with a high confidence are used (clusters with more than 10 reads or individual unique sequences with more than five reads). For targets (such as RBD) with fewer initial sequences, the threshold of individual unique sequences was set to 1. The final selected sequences were then randomly divided into training and test groups in an 8:2 ratio to construct the training and test sets. The statistical information of the data against three targets used in this study is listed in Supplementary Table 1.

GRAPE-LM architecture

GRAPE-LM is a deep learning model that leverages a CAE powered by a nucleic acid LM to evolve RNA aptamer sequences. GRAPE-LM consists of three key components: (1) an encoder that maps RNA sequences to a continuous latent semantic space; (2) an activity guidance conditional module that constrains this latent space (Extended Data Fig. 1a), enabling it to organize aptamer sequences according to their pseudo-activity levels (Extended Data Fig. 1b); and (3) a decoder that generates potential RNA aptamer sequences by sampling from potential high-activity regions of the latent space.

Unlike traditional VAEs with Kullback–Leibler divergence regularization, we regularize the latent space with task-aligned constraints: (1) a constraint on sequence reconstruction ensures that the model can consistently generate meaningful RNA sequences and (2) an activity alignment loss pulls embeddings of higher-activity sequences toward target-specific prototypes, which yields diversity around activity-relevant directions while maintaining the capability for reconstruction by the RNA-LM-powered decoder.

The CAE is also an autoregressive generation framework. In the decoder, we employ an autoregressive generation mode, using latent space feature vectors as starting symbols to predict the first base of RNA sequences and then predict the $n + 1^{\text{th}}$ base according to the first n bases.

LM-based encoder. LMs originally developed for natural language processing have found promising applications in biomolecule engineering, such as protein design^{64–66} and evolution^{5,67,68}. In parallel, nucleic acid-focused LMs have also been explored recently⁶⁹. Current research predominantly focuses on predictive tasks, leaving a gap in leveraging LMs for de novo RNA sequence design or evolution—an area where protein design has already seen notable success^{5,67}. Bridging this gap could unlock new opportunities for RNA-based therapeutics development and synthetic biology applications.

To improve generalization and reduce sensitivity to individual data points, GRAPE-LM's encoder combines a nucleic acid LM with

a trainable adapter. The adapter, implemented as a fully connected neural network with batch normalization, maps LM-derived features into a 128-dimensional latent space. It can mitigate potential biases from the LM while tailoring the encoder to aptamer-specific tasks. Six LMs were evaluated in this study: RNA-FM²⁶, RNA-BERT²⁷, RNAErnie²⁸, EvoFlow-RNA²⁹, RiNALMo³⁰ and Evo³¹. These RNA LMs are BERT-style transformer-based RNA foundation models pretrained on large-scale RNA sequences (Supplementary Fig. 1a); Evo³¹ is a StripedHyena-based genomic foundation model trained on 2.7 million prokaryotic and phage genomes (Supplementary Fig. 1b). RNA LMs were selected for their RNA-specific focus, whereas Evo's broader sequence compatibility (DNA, RNA and proteins) offers additional flexibility. During the training phase, we freeze the weights of the LMs and only update the adapter.

Transformer-based decoder. Given the similarity between RNA sequence generation and text generation, we designed the RNA aptamer decoder based on the transformer network⁷⁰, which generates aptamer candidates from sampled points in the latent space. The transformer network offers advantages in generative tasks^{71,72} due to its flexible self-attention mechanism, efficient parallel computation and strong long-range dependency capture capabilities. Unlike traditional sequence-to-sequence models, GRAPE-LM's decoder employs a self-attention mechanism to iteratively generate nucleotides without relying on fixed-length context vectors. The self-attention mechanism allows each generated nucleotide to be dynamically weighted based on all previously generated nucleotides, enabling the decoder to consider context at each step and generate more rational RNA sequences. To enhance coherence and avoid limitations of bidirectional attention in the encoder–decoder framework, a decoder-only structure was adopted. The decoder includes two transformer–decoder layers, each with two attention heads and a hidden dimension of 128. The dropout technique with a rate of 0.05 is applied between layers to prevent overfitting.

Activity guidance module. This module constrains the latent space by predicting pseudo-activity scores (Extended Data Fig. 1b) through a multi-layer regression model. It consists of two layers of feedforward neural networks, with a hidden layer containing 128 neurons activated by the ReLU function and an output layer consisting of a neuron activated by the sigmoid function. This module's input is the latent space feature vector of a sequence (the encoder output of the corresponding sequence) and outputs the sequence's predicted pseudo-activity score. The dropout technique was adopted at a rate of 0.05 to ensure robustness against overfitting.

Pseudo-activity is derived from post-enrichment sequence distributions and serves as an indirect measure of biological activity. Although these distributions encapsulate the overall results of a screening round, they may be influenced by technical biases, such as PCR amplification artifacts, procedural variability and fluctuations in transfection efficiency and cell clonal viability. Experience with both multi-round CRISmers and conventional SELEX shows that library diversity changes substantially over successive rounds of selection^{11,73}. Nevertheless, with only one round of screening data available, this distribution remains the sole actionable input for AI-driven analysis. To reduce the impact of these noises on the model, we aggregate read counts across multiple sequences within a cluster to compute the pseudo-activity score. This aggregation reduces stochastic variability and offsets biases that would be amplified if only single-sequence counts were considered. Central to this study is the assumption that differential enrichment, defined as increased abundance of aptamer populations in a specific CRISmers round, directly correlates with the intrinsic binding affinity and targeting selectivity of aptamers. The pseudo-activity score serves as a practical high-throughput measurement of relative 'hitness' of each aptamer sequences within a round of the CRISmers.

Training and evaluation

Loss function. The GRAPE-LM model optimizes two objectives: (1) accurate sequence reconstruction and (2) discriminative latent space organization. These are quantified via a weighted fusion of two losses: reconstruction loss and activity prediction loss. Reconstruction loss calculates the cross-entropy (CE) between input and generated sequences, and activity prediction loss calculates the mean squared error (MSE) between predicted pseudo-activity scores and corresponding ground truth. The total loss is defined as

$$\text{Loss} = w \times \text{MSE}(y, \hat{y}) + (1 - w) \times \text{CE}(s, \hat{s})$$

where s and y denote the input sequence and its pseudo-activity score, and $\hat{s}$ and $\hat{y}$ are their reconstructed counterparts. The weight w balances the two objectives: $w = 0.5$ for the CD3ε and c-Myc targets and $w = 0.85$ for RBD.

Training protocol. The hyperparameters used in this study were optimized using grid search (Supplementary Note 1) while considering computational efficiency and model performance (Supplementary Fig. 12). Training of GRAPE-LM was conducted on a server with one AMD EPYC 7642 GPU with 512 GB RAM memory and one NVIDIA RTX 4090 GPU with 24 GB memory using the Adam optimizer with dynamically adjusting the learning rate. Models were trained for up to 400 epochs, with early stopping triggered if test loss plateaued.

Aptamer candidate generation. We used the unique sequences obtained from the pre-GRAPE-LM CRISmers screening as a seed library. For each iteration, n ($n \geq 2$) seeds were randomly selected and mapped to the latent space via the GRAPE-LM encoder. The midpoint of these n seeds in the latent space served as the sampling point for generating novel sequences by the GRAPE-LM decoder. Notably, GRAPE-LM's latent space can cluster RNA sequences according to their pseudo-activity levels, enabling directed sampling from potential high-activity regions by using top active seeds—a capability absent in traditional sequence similarity-based latent spaces.

In the present study, we used the top 2,000 unique sequences by read count as top active seeds. Although increasing the number of selected seeds (n) theoretically expands combinatorial possibilities for novel sequence generation, $n = 2$ proved sufficient to meet the required sample generation scale for this investigation. Consequently, all experiments employed $n = 2$ for sampling. Gaussian noise was systematically incorporated into the latent space vectors during the sampling process to enhance sample diversity. This methodological framework ensures directional exploration of activity-specific regions while maintaining structural novelty in generated sequences.

Evaluation metrics. Computational performance was assessed using four metrics:

Recovery rate. A quantitative measure of a generative model's effectiveness, calculated by comparing the generated aptamer candidate sequence to the original one and finding the fraction of correctly recovered nucleotides.

Diversity. Evaluates the ability of a generative model to generate large-scale sequences by comprehensively considering the internal diversity ($\text{Diversity}_{\text{inter}}$) and novelty of the generated sequences. It is calculated with the following formulas:

$$\text{Diversity}_{\text{inter}} = \frac{N_{\text{cluster}}}{N}$$

$$\text{Novelty} = 1 - \frac{N_{\text{hit}}}{N}$$

$$\text{Diversity} = \frac{2 \times \text{Diversity}_{\text{inter}} \times \text{Novelty}}{\text{Diversity}_{\text{inter}} + \text{Novelty}}$$

where N_{cluster} is the number of clusters among generated sequences clustering by using the Cd-hit program with a threshold of 0.8; N is the number of generated sequences (10,000 used in this study); and N_{hit} is the number of the generated sequences that share equal or more than 80% sequence similarity (calculated by Cd-hit with a threshold of 0.8) with pre-AI top 10,000 leads.

GC content. Measures GC deviation from a reference pool (CRISmers screened aptamers), therefore providing a coarse but informative measure of whether a generative model produces sequences whose global compositional properties remain consistent with experimentally enriched aptamer families.

Pearson's correlation coefficient. Quantifies the effectiveness of activity guidance by measuring the correlation of predicted pseudo-activity scores and the pseudo-activity scores calculated from the output of CRISmers.

Recovery rate is defined as the reconstruction quality of experimentally validated aptamers in the generated library, serving as a proxy for functional recall, and indicates the ability of the model to regenerate sequences with known binding activity. Diversity reflects the potential to explore a vast sequence space, which is critical for avoiding local optima and increases the likelihood of identifying novel high-affinity binders. GC content reflects key biophysical/biochemical properties such as thermodynamic stability and secondary structure formation, which influence RNA folding and recognition^{4,74,75}. Previous SELEX studies^{76–78} show that enriched aptamer pools often converge toward target-specific nucleotide composition constraints, so assessing GC content deviation provides a model-agnostic indicator of whether generated sequence pools preserve biologically reasonable global composition. Although these metrics do not measure binding affinity directly, they together provide heuristic proxies for functional aptamer discovery and experimental success.

CRISmers screening and validation

Construction of plasmids of CRISmers library. The human CD3ε extracellular domain (amino acids D23–D126, UniProt no. P07766), codon-optimized RBD of SARS-CoV-2 spike protein (S1, amino acids 319–541, GISAID ID: EPI_ISL_414631) and c-Myc (UniProt no. 01106) were commercially synthesized by GenScript (Supplementary Note 2). The CD3ε, RBD or c-Myc fragment was subsequently cloned into the pHU6-gLuc sgRNA-scaffold-1.2 BsmBI-scaffold-polyT-E F1a-NLS-linker-EcoRI-EcoRI-linker-VP64-p65-HSF1 (VPH) vector at the EcoRI sites. AI-designed 20-bp oligonucleotide pools with 90-bp overlapping sequences (synthesized by Twist Bioscience; Supplementary Note 3) underwent limited-cycle PCR amplification (15–20 cycles) using Phusion Hot Start Flex 2× Master Mix (New England Biolabs (NEB), M0536L) to minimize amplification bias. The resulting approximately 200-bp products were column purified using the QIAquick PCR Purification Kit (Qiagen, 28104) and assembled into BsmBI-digested vectors via Gibson Assembly. After isopropanol precipitation, the ligation products were transformed into electrocompetent *Escherichia coli* cells (NEB, 10β) and plated onto LB agar plates supplemented with appropriate antibiotics. The plates were then incubated at 32 °C for 18 hours. Library quality control demonstrated 99.9% sequence coverage with an average of 500× per clone, as determined by Maxiprep plasmid quantification and NGS validation. Sequencing for the constructed screening library was conducted by Yomebio Co., Ltd.

Lentiviral CRISmers library production. HEK293T cells (American Type Culture Collection (ATCC), CRL-3216) were cultured in DMEM (Gibco, 11965092) supplemented with 10% FBS (ExCell, FSP500) at

37 °C/5% CO₂. Upon reaching 80–90% confluency in 10-cm culture dishes, cells were co-transfected with the CRISmers library plasmid, psPAX2 packaging plasmid (Addgene, 12260) and pMD2.G envelope plasmid (Addgene, 12259) using polyethyleneimine (PEI; Polysciences, 23966-1) at 3:1 PEI:DNA ratio. The transfection medium was replaced with fresh complete medium 6 hours after transfection. Viral supernatants were collected at 48 hours and 72 hours and filtered through 0.45-μm syringe filters (Pall, 4654). Aliquots were stored at –80 °C.

Lentiviral CRISmers library transduction. The CRISmers library was transduced into monoclonal 4-53 dCas9-gLuc-puro reporter cells (established per ref. 11). In brief, suspended cells were incubated with lentiviral particles in medium containing 8 μg ml^{–1} polybrene (Sigma-Aldrich, TR-1003-G). After 8-hour incubation at 37 °C/5% CO₂, the medium was replaced with fresh complete medium. The multiplicity of infection was maintained at 0.1–0.2 to ensure single-copy integration, with library coverage exceeding 500 cells per aptamer sequence.

CRISmers selection. The transduced cells were seeded at a low density and incubated for an additional 48 hours. Subsequently, puromycin (InvivoGen, ant-pr-1) was added to the culture medium at a concentration of 2 μg ml^{–1} to select for drug-resistant cells. Medium replacement with drug-containing fresh medium was performed every 48 hours until complete death of untransduced controls. After 6–8 days of expansion, genomic DNA was extracted. Aptamer sequences were amplified with High-Fidelity Polymerase Phusion Hot Start Flex 2× Master Mix (NEB, M0536S) using specific primers (forward: 5′-TATGTTTAAGAGCTAGAAATAGCAAGTTAAAATAAGGCTA-3′; reverse: 5′-CCATCTTTGCAAAGCTTATATCGATAAAAAGCACCGACTC-3′) and subjected to NGS analysis (GENEWIZ).

NGS analyses. The PCR products were subjected to sequencing on an Illumina NovaSeq 6000 platform using 150-bp paired-end sequencing. For the pre-AI CD3ε screening, a total of 6,949,288 raw reads were obtained. The pre-AI RBD screening yielded 3,943,669 raw reads. The pre-AI screening on c-Myc yielded 6,729,453 raw reads. The read counts of each sequence were ranked and analyzed in comparison, as discussed in this study. Selected sequences were then further validated through additional experiments.

Luciferase reporter assay. Selected CD3ε-targeting, RBD-targeting or c-Myc-targeting aptamers were further validated using a CRISPR-Cas-based luciferase reporter assay (Fig. 3a). These aptamers were synthesized by Sangon Biotech and then cloned, via the BsmBI sites, into the vector pHU6-gLuc sgRNA-scaffold-1.2 BsmBI-scaffold-polyT-E F1a-NLS-linker-EcoRI-CD3ε ED/RBD/c-Myc-EcoRI-linker-VPH (Supplementary Note 2). HEK293T cells were seeded at 2.5 × 10⁴ cells per well in 96-well plates, and transient transfection was performed the following day using PEI (Polysciences, 24885-2) at a 3:1 PEI:DNA ratio. Every well was co-transfected with 20 ng of 9× gLuc target *Gaussia* luciferase plasmid, 85 ng of dCas9 expression plasmid, 85 ng of hU6-gLuc sgRNA-scaffold-1.2 aptamer-scaffold-polyT-E F1a-NLS-linker-EcoRI-CD3ε ED/RBD/c-Myc-EcoRI-linker-VPH plasmid and 10 ng of firefly expression plasmid, respectively. The pHU6-gLuc sgRNA-scaffold-1.2 BsmBI-scaffold-polyT-E F1a-NLS-linker-EcoRI-CD3ε ED/RBD/c-Myc-EcoRI-linker-VPH plasmid was used as a negative control. Six hours after transfection, the medium was replaced with fresh medium. Luminescence was quantified 48 hours after transfection using *Gaussia*/firefly luciferase assay kits (Vigorous Biotechnology, T004/T003) with a luminescence counter (PerkinElmer).

Generation of CD3ε-knockout Jurkat cell lines using CRISPR-Cas9. CD3ε-knockout Jurkat cell lines were generated using CRISPR-Cas9-mediated genome editing. A plasmid co-expressing SpCas9–

GFP and a CD3ε-specific sgRNA (target sequence: 5′-GATGTCCACT ATGACAATTG-3′) were transfected into Jurkat cells (clone E6-1; ATCC, TIB-152) using the ProteanFect transfection kit (Nanoportal Biotech, PT01) according to the manufacturer's instructions. Cells were stained with PE anti-human CD3 antibody (BioLegend, 317308; clone, OKT3; isotype PE mouse IgG2a, κ isotype Ctrl) and subjected to fluorescence-activated cell sorting (FACS). Single cells that were GFP positive and PE negative were sorted into 96-well plates for clonal expansion.

After expansion, monoclonal populations were initially screened for CD3ε surface expression by flow cytometry. Clones exhibiting stained fluorescence levels similar to those of the PE isotype negative control were selected for genotypic validation. Genomic DNA was extracted from candidate clones, and the region flanking the CRISPR target site was PCR amplified and analyzed by Sanger sequencing using the primers 5′-CCACGGCCACCACTCTTTGCTT-3′ (forward) and 5′-GGGCGGAATCGGATGACTGCAG-3′ (reverse). Clones showing double peaks in the sequencing chromatogram, indicative of insertion and deletion (indel) mutations, were further validated. The PCR products were cloned into a blank vector, and multiple bacterial colonies were Sanger sequenced to accurately determine indel patterns and confirm biallelic mutations. For cross-validation, subsequent examinations of aptamer specificity were performed on two independent clones.

Flow cytometry. 5′-Cy5-labeled 2′-F-*O*-modified RNA aptamers were synthesized by General Biol containing 2′-fluorine pyrimidines (C/U) and 2′-*O*-methyl purines (A/G), and were reconstituted in nuclease-free distilled water to a 10 μM stock. Harvested cells were centrifuged at 1,000g for 5 minutes, washed with PBS and resuspended in PBS. Jurkat cells (5 × 10⁵ per tube) were incubated with 100 nM aptamers in a 200-μl final volume for 20 minutes at 25 °C. After three PBS washes, cells were resuspended in 100 μl of PBS and analyzed immediately on a CytoFLEXS (Beckman Coulter). Cells were gated using forward scatter area (FSC-A) versus side scatter area (SSC-A) parameters for debris exclusion, followed by forward scatter height (FSC-H)/FSC-A pulse processing for singlet discrimination. Fluorescence thresholds in the APC channel (detection channel for Cy5) were defined by unstained cell populations.

For flow cytometry-based binding affinity measurements, the aptamers were identically modified (5′-Cy5-labeled, 2′-F-pyrimidine, 2′-*O*-methyl-purine RNA). Aptamers were serially diluted by threefold with an initial concentration of 1,000 nM to generate eight working concentrations, with one blank control (no aptamer) included for a total of 10 conditions. Jurkat cells (5 × 10⁵ per tube) were incubated in a final volume of 200 μl for 20 minutes at 25 °C. Subsequent steps—washing, resuspension and flow cytometric analysis—were performed as above, using the same instrument and gating strategy.

MST. SARS-CoV-2 spike protein (Sino Biological, 40589-V08B1) was diluted to 9.3 μM in nuclease-free distilled water. SARS-CoV-2 spike S2 protein (Sino Biological, 40590-V08B) was diluted to 10.528 μM in nuclease-free distilled water. Recombinant human c-Myc active protein was obtained from Abcam (0.5 mg ml^{–1}; ab169901). 5′-Cy5-labeled 2′-F-*O*-modified RNA aptamers were synthesized and diluted to 20 nM in TE buffer. The experimental procedures were performed as described in a previous study⁴⁶. In brief, 16 200-μl RNase/DNase-free and low-binding PCR tubes were prepared for serial dilution (Axygen, PCR-02-L-C). Proteins (spike: 4.65 μM, S2: 5.26 μM, c-Myc: 4.72 μM) underwent a twofold serial dilution across 16 tubes, each preloaded with 10 nM of the matching aptamer, followed by a 15-minute incubation at 25 °C. Monolith capillaries (NanoTemper, MO-K022) were loaded with each sample and positioned in the MonolithX instrument (NanoTemper, MM-233). Excitation parameters were optimized through a pre-test, with the infrared laser power set to medium. *K_d* values were calculated using MO.Affinity Analysis software version 2.3 (NanoTemper).

Pseudovirus neutralization assay. HEK293T cells were plated in six-well culture plates and maintained under standard conditions until reaching 80–90% confluency prior to transfection. PEI-mediated transfection was performed using the following plasmid combination per well: 2 µg of cytomegalovirus (CMV)-driven luciferase reporter construct and 0.4 µg of either spike (wild-type) or other variant expression vectors, supplemented with 0.2 µg each of TAT, Rev and GAGpol accessory plasmids. After a six-hour incubation period after transfection, the transfection medium was replaced with fresh DMEM supplemented with 10% FBS. Viral particles were subsequently collected at 48-hour and 72-hour timepoints through sterile filtration using 0.45-µm Pall syringe filters.

HEK293T-hACE2 stable cells were plated in 96-well plates at a density of 2×10^4 cells per well and cultured overnight under standard conditions (37 °C, 5% CO₂). Serial dilutions of cholesterol-PEG40K-2'-F-O-modified aptamers were prepared in pseudovirus and incubated for 1 hour at room temperature. The pretreated pseudo virus-aptamer mixtures were then transferred to the cell and further incubated for 2 hours. After the viral challenge, the supernatant was removed by aspiration and replaced with complete fresh medium. The cells were then incubated for an additional 48 hours at 37 °C in a 5% (v/v) CO₂ incubator. Luciferase activities were measured sequentially using a luminescence counter (PerkinElmer).

Figure preparation

Figures were prepared with GraphPad Prism, PowerPoint, Adobe Photoshop, Adobe Illustrator and BioRender.com. Icons in figures were created with BioRender.com with the following specific citations: <https://BioRender.com/8cn0ckx>, <https://BioRender.com/83bi9am> and <https://BioRender.com/w85g588>.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The CRISmers sequencing data from both primary screen and pooled examination are available from Zenodo: <https://zenodo.org/records/18050896> (ref. 11) and <https://doi.org/10.5281/zenodo.18005327> (ref. 79). The minimum datasets used and model checkpoints for GRAPE-LM are available from GitHub: <https://github.com/tansaox2008123/GRAPE-LM>. Source data are provided with this paper.

Code availability

The source code of GRAPE-LM is available at <https://github.com/tansaox2008123/GRAPE-LM>. Through the GRAPE-LM online platform (<https://grape-lm.bioailab.net/>), researchers can easily retrieve aptamer sequences designed for three specific molecular targets featured in this publication.

References

- Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
- Ni, B., Kaplan, D. L. & Buehler, M. J. ForceGen: end-to-end de novo protein generation based on nonlinear mechanical unfolding responses using a language diffusion model. *Sci. Adv.* **10**, ead14000 (2024).
- Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
- Zhang, Q. et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nat. Commun.* **16**, 1553 (2025).
- Singh, A. Protein language models guide directed antibody evolution. *Nat. Methods* **20**, 785 (2023).
- Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
- Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates, 2017).
- Castro, E. et al. Transformer-based protein generation with regularized latent space optimization. *Nat. Mach. Intell.* **4**, 840–851 (2022).
- Ferruz, N. & Höcker, B. Controllable protein design with language models. *Nat. Mach. Intell.* **4**, 521–532 (2022).
- Nguyen Quang, N., Bouvier, C., Henriques, A., Lelandais, B. & Ducongé, F. Time-lapse imaging of molecular evolution by high-throughput sequencing. *Nucleic Acids Res.* **46**, 7480–7494 (2018).
- Chan, C. Y. et al. A structural interpretation of the effect of GC-content on efficiency of RNA interference. *BMC Bioinformatics* **10**, S33 (2009).
- Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999).
- Molejon, N. A. et al. Selection of G-rich ssDNA aptamers for the detection of enterotoxins of the cholera toxin family. *Anal. Biochem.* **669**, 115118 (2023).
- Melikishvili, M. et al. SELEX identifies high-affinity RNA targets for chromatin-binding proteins PARP1 and MeCP2. *iScience* **28**, 113299 (2025).
- Thiel, W. H. et al. Nucleotide bias observed with a short SELEX RNA aptamer library. *Nucleic Acid Ther.* **21**, 253–263 (2011).
- Zhang, J. et al. CRISmers NGS data used and generated by GRAPE-LM (1.0). *Zenodo* <https://doi.org/10.5281/zenodo.18005327> (2025).

Acknowledgements

We thank all colleagues in our laboratories for experimental assistance and helpful discussions. This work was supported by the National Key Research and Development Program of China (no. 2023YFA0915000 to Y.W.), the National Natural Science Foundation of China (no. 91957121 and no. 82273967 to Y.W., no. 82273890 to Y.Z. and no. 62302311 to Jun Zhang), the Department of Science and Technology of Guangdong Province (no. 2021QN020576 to Y.W.), the Guangdong Basic and Applied Basic Research Foundation (no. 2024A1515011681 to Jun Zhang), the Shenzhen Science and Technology Program (no. ZDSYS20220303153551001 to Y.W.), the Shenzhen Stable Support Grant (no. GXWD 20231130103401001 to Y.Z.), the Shenzhen Science and Technology Program (no. JCYJ20240813104817024 to Y.Z.), the China Postdoctoral Science Foundation (no. 2023M742397 and no. 2024T170585 to Ju Zhang), the Postdoctoral Fellowship Program of the China Postdoctoral Science Foundation (no. GZC20231724 to Ju Zhang) and the Internal Fund of the National Engineering Laboratory for Big Data System Computing Technology (no. SZU-BDSC-IF2024-01 to Jun Zhang).

Author contributions

Y.W., Y.Z. and Jun Zhang conceived and designed the study. Jun Zhang, S.T., H.Z. and X.M. performed modeling and computational analyses. Ju Zhang, C.L., Y.C. and B.L. performed wet lab experiments. All authors analyzed the data and interpreted the results. Y.W., Y.Z., Jun Zhang and Ju Zhang wrote the manuscript, with contributions from all coauthors. Jun Zhang and Ju Zhang contributed equally to this work. Y.W. and Y.Z. supervised the project. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

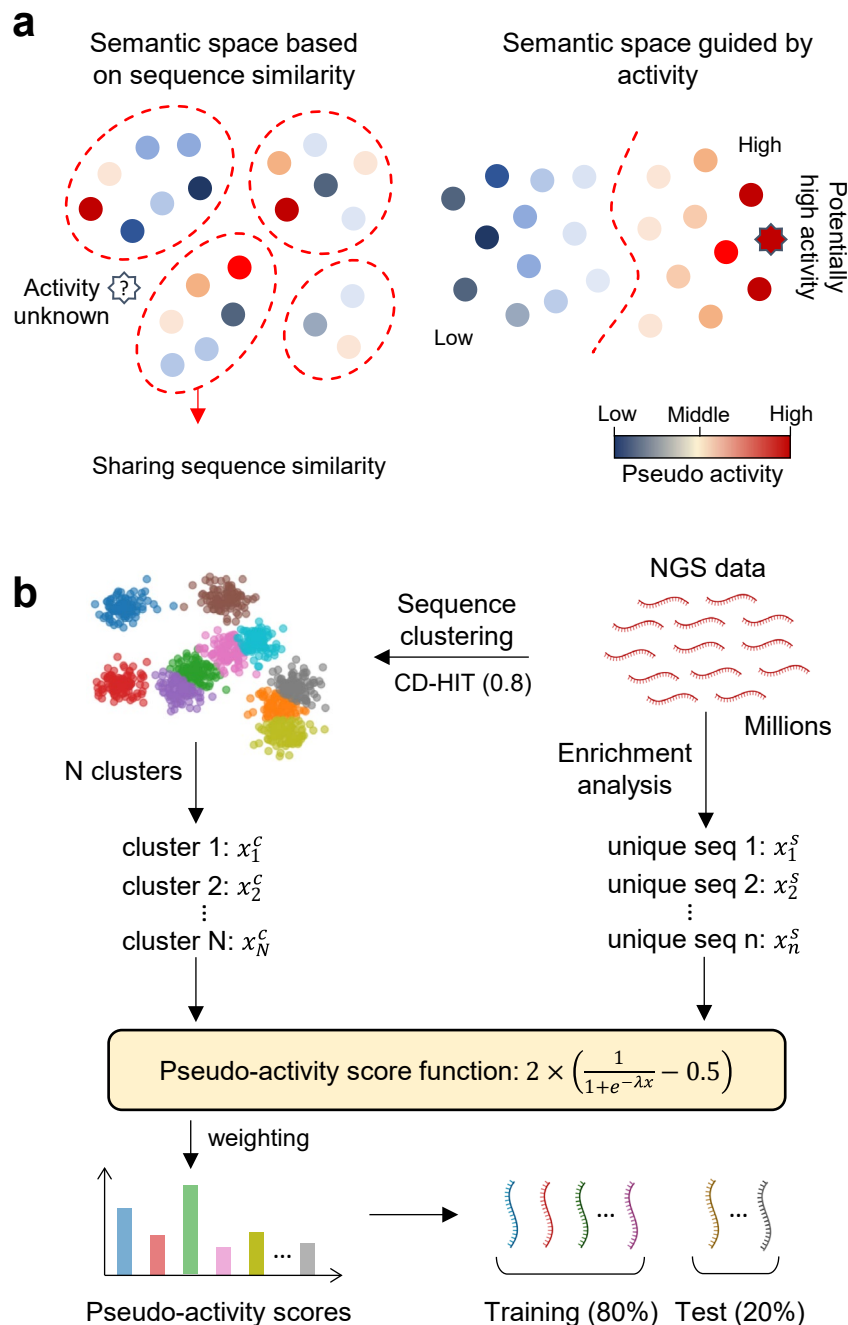
Extended data is available for this paper at <https://doi.org/10.1038/s41587-026-03007-5>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-026-03007-5>.

Correspondence and requests for materials should be addressed to Jun Zhang, Yang Zhang or Yu Wang.

Peer review information *Nature Biotechnology* thanks Pranam Chatterjee and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

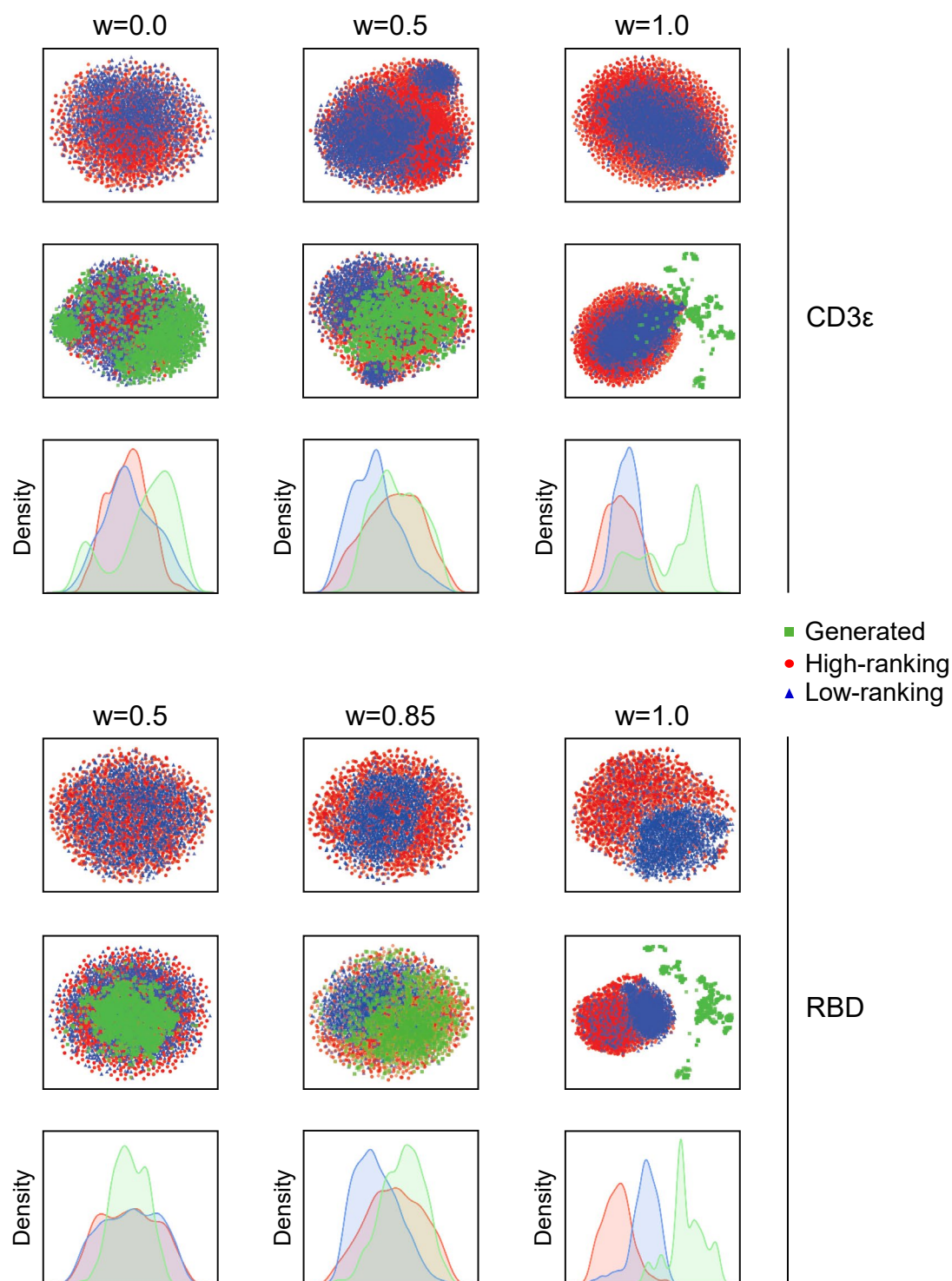
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Comparison of two latent spaces and data processing.

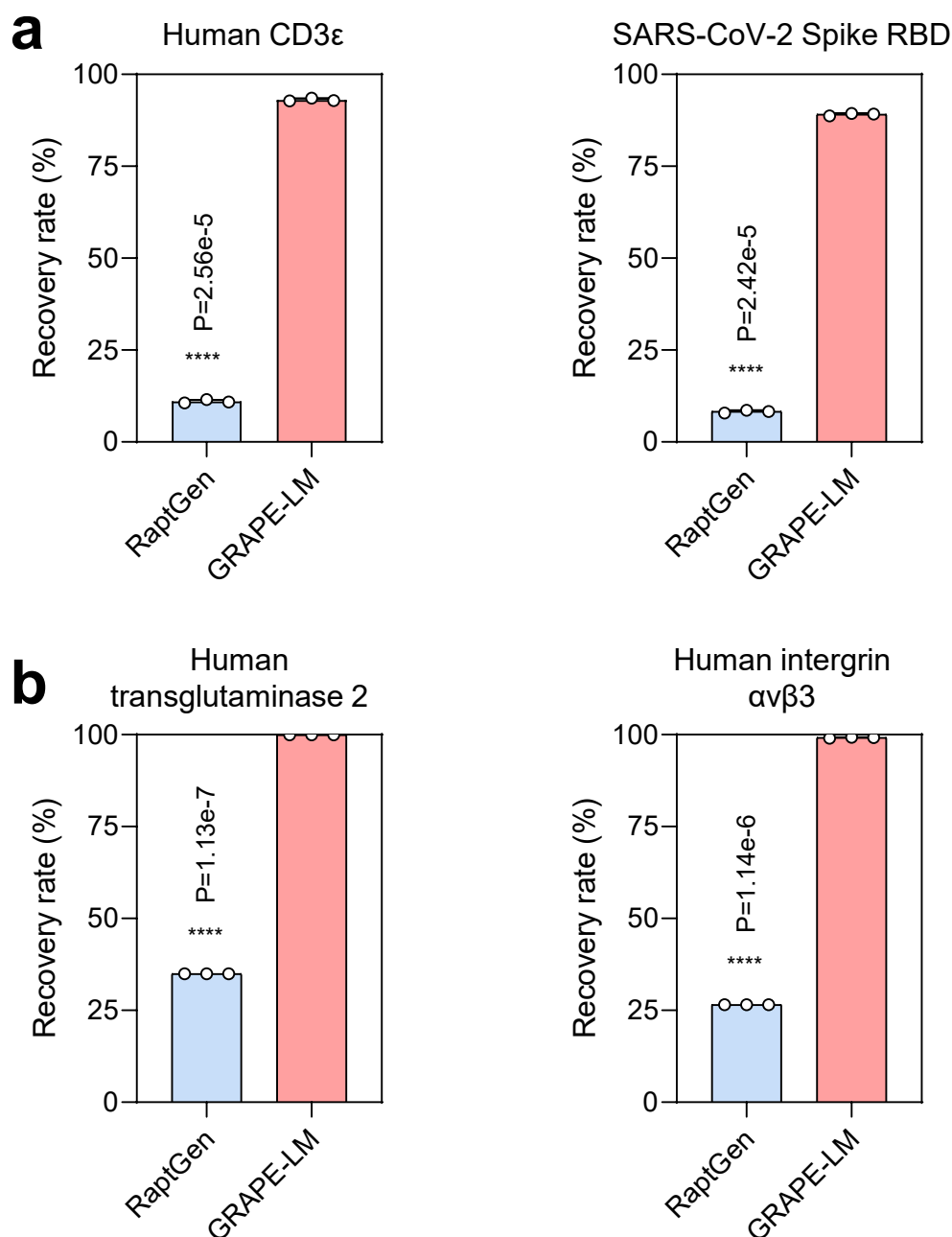
(a) Comparison of activity-guided versus sequence similarity-based semantic spaces. The activity-guided latent space (constrained by pseudo-activity scores) enables focused sampling of functional aptamers, while traditional sequence similarity-based spaces permit only random exploration. (b) Data processing pipeline. Training and test datasets were derived from the first-round output of CRISmers (an intracellular CRISPR/Cas-based aptamer screening system). Unique sequences were assigned pseudo-activity scores (0–1 scale) based on enrichment frequencies. λ is a hyperparameter that needs to be optimized for

different targets, and it is recommended to initially use 0.05. The sequences were clustered using CD-HIT with a threshold of 0.8. Sequences only with a high confidence were used (clusters with more than 10 reads or individual unique sequences with more than 5 reads). For targets (such as RBD) with fewer initial sequences, the threshold of individual unique sequences was set to 1. The final selected sequences were then randomly divided into training and test groups in an 8:2 ratio to construct the training and test sets. Icons in figures were created with BioRender (<https://www.biorender.com/>).

**Extended Data Fig. 2 | Exploring the role of activity guidance strategy.**

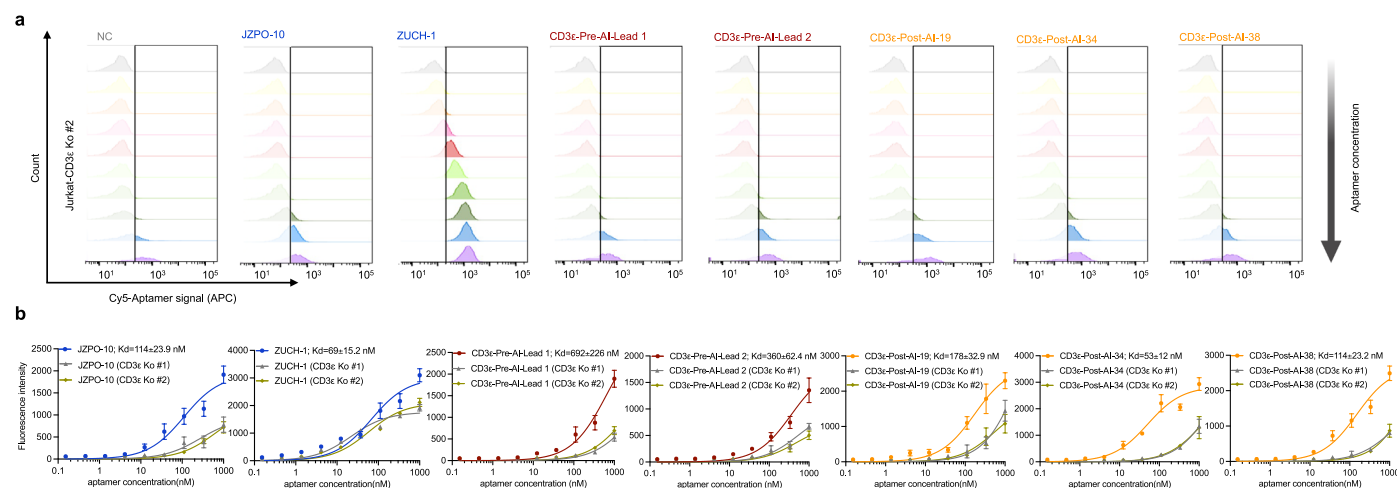
To evaluate the regulatory role of the activity-guidance module on latent space organization, we analyzed how varying the activity-guidance loss weight influenced the separation between high-ranking and low-ranking activity

samples. Using t-SNE, we mapped the latent features generated by GRAPE-LM's encoder of 2,000 samples per group from high-dimensional to 2D space, facilitating cluster analysis of generated aptamers based on functional similarity.



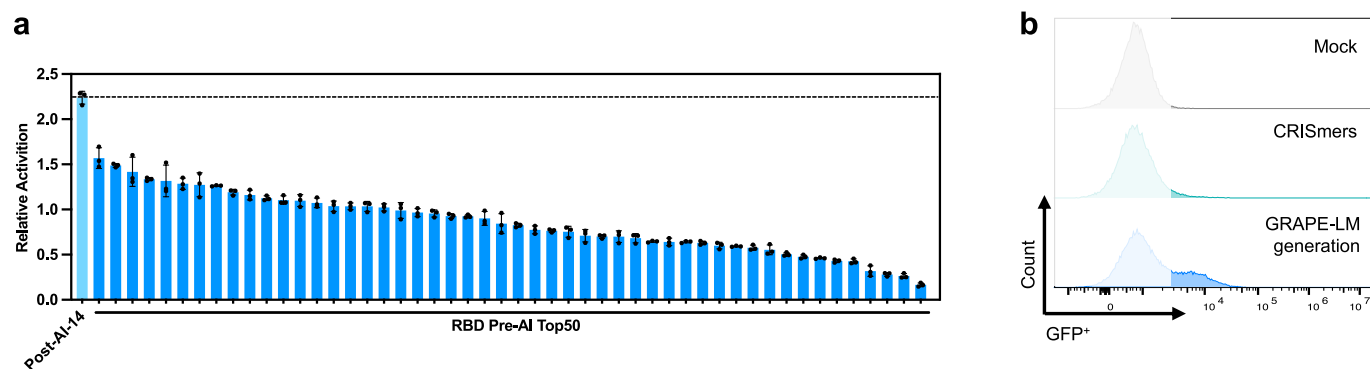
Extended Data Fig. 3 | Comparative results of GRAPE-LM and RaptGen in terms of the recovery rate. The overall recovery rates are calculated using the corresponding model on the test sets of four targets from CRISmers (a) and

SELEX (b). Data plot mean with standard deviation (SD), $n = 3$ experimental replicates. The statistical analyses were conducted using a two-sided Student's t-test (**** $P < 0.0001$).



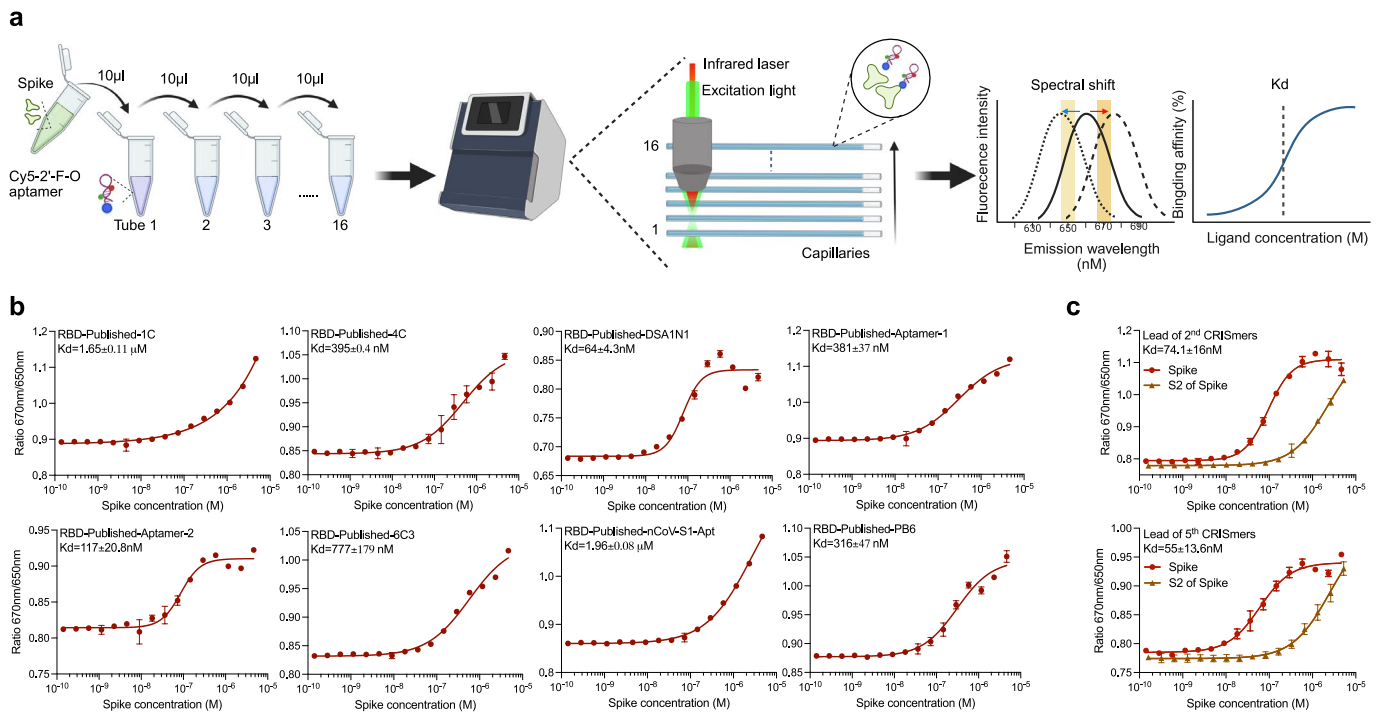
Extended Data Fig. 4 | Dose-dependent flow cytometry analyses of the Cy5-labeled CD3ε aptamers. (a) Representative flow cytometry histograms showing fluorescence intensity shifts in CD3ε-Ko #2 cell after treatment with serially diluted Cy5-aptamers. **(b)** Dose-response binding curves of the Cy5-labeled CD3ε aptamers on Jurkat cells. CD3ε knock-out cells (Jurkat CD3ε-Ko

#1 and #2) serve as negative control to assess binding specificity. Apparent equilibrium dissociation constants (K_d) were derived from nonlinear regression analysis (one-site specific binding model) using GraphPad Prism. Data represent mean $\pm$ SD ($n = 3$ biological replicates).

**Extended Data Fig. 5 | Characterization of aptamer binding affinity for CD3 ϵ .**

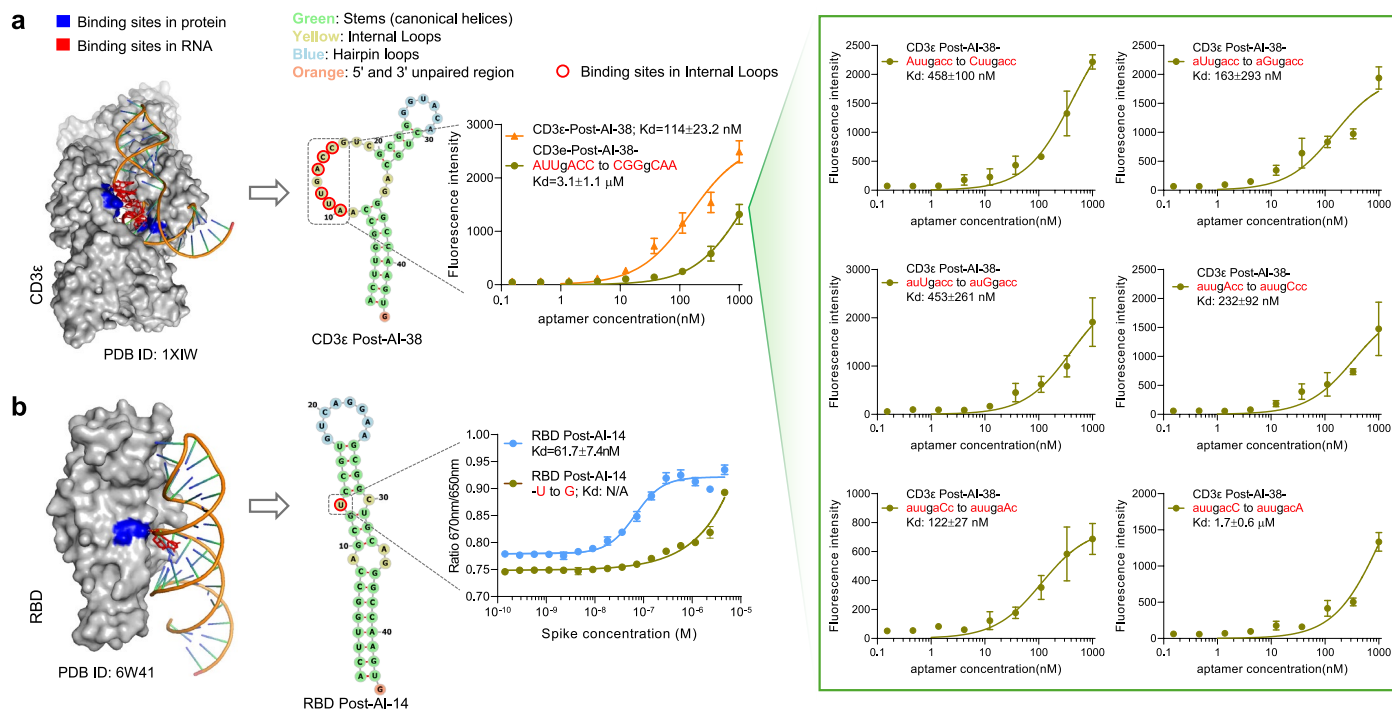
(a) Luciferase reporter assay results of the top 50 candidate sequences from one round of CRISmers screening, ranked by sequencing read abundance. Data represent mean with SD, $n = 3$ biological replicates. (b) Comparative GFP reporter assay of libraries from one-round screening using CRISmers versus GRAPE-LM.

“Mock” denotes transfection with unrelated control plasmids; “CRISmers” represents transfection with the sub-library constructed after one round of CRISmers screening; “GRAPE-LM” corresponds to transfection with the library generated by GRAPE-LM in a single round.



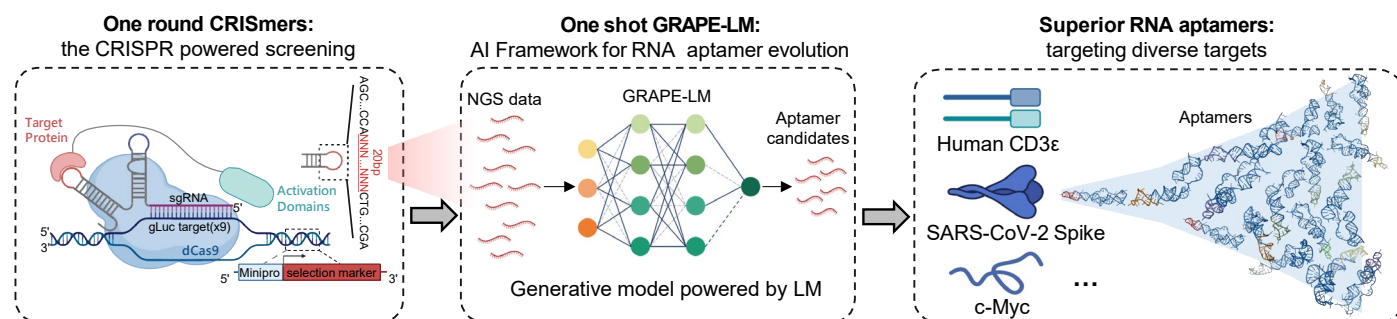
Extended Data Fig. 6 | The Microscale Thermophoresis examination and the additional results for RBD. (a) The schematic diagram of the Microscale Thermophoresis (MST). Icons in figures were created with BioRender (<https://www.biorender.com/>). (b) MST binding curves of previously reported SELEX-derived aptamers targeting the RBD of SARS-CoV-2 Spike protein. Data showed mean with SD, with $n = 3$ biological replicates. (c) Results of Kd value determination based on MST for the Lead of 2nd round and the Lead of 5th

round from iterative CRISMsers. Data showed mean with SD, with $n = 3$ biological replicates. The receptor-binding domain (RBD) resides in the S1 subunit of the SARS-CoV-2 Spike glycoprotein. To validate binding specificity during MST assays, the S2 subunit—a non-target structural region—was used as a negative control, ensuring the aptamer selectively recognizes the RBD rather than unrelated epitopes.



Extended Data Fig. 7 | Case study of the internal loop motif in two representative aptamers. Prediction of secondary structures and functional analysis of predicted binding sites for representative CD3ε (**a**) and RBD

(**b**) aptamers. The dashed box highlights the computationally predicted binding site within the internal loop. Data represent the mean with SD (n = 3 biological replicates).



Extended Data Fig. 8 | A powerful new paradigm of accelerated RNA aptamer evolution. This new paradigm is enabled by one shot GRAPE-LM introduced in this work, informed by one round CRISmers. Icons in figures were created with BioRender (<https://www.biorender.com/>).

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection We used CD-Hit (version 4.6.7) for sequence clustering.

Data analysis We used python custom scripts (python 3.8.11 and pytorch 2.4.1) for data analysis, AlphaFold 3 and VFold for RNA 3D structure prediction, RNAFold for RNA secondary structure prediction. For sequence and structure alignments, we used Clustal Omega and US-align, respectively. We used WebLogo 3 for the visualization of sequence patterns, HDOCK for the molecular docking, and PyMOL (version 1.8) for the illustration of the 3D structures. Figures are generated with GraphPad Prism (version 8.0.2), PowerPoint, Adobe Photoshop, Adobe Illustrator, and BioRender.com. All custom scripts are provided on GitHub (<https://github.com/tansaox2008123/GRAPE-LM>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The numerical data supporting the findings of this paper are provided in the Source Data, and the sequences can be generated by running GRAPE-LM. The datasets used and model checkpoints for GRAPE-LM are available from Github: <https://github.com/tansaox2008123/GRAPE-LM>. The sequencing data is deposited on Zenodo: <https://doi.org/10.5281/zenodo.18005327>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of biological replicates and sample sizes are indicated in the figure legends
Data exclusions	All data points are shown without exclusion of outliers.
Replication	Wet-lab experiments involving molecular and cellular assays were conducted with three biological replicates.
Randomization	The order of training data was randomized prior to model training. In CRISmers screening and validation, RNA aptamers were randomly synthesized and sampled. For the cellular experiments, cells were plated at uniform density and randomly allocated to treatment groups prior to experimentation.
Blinding	The identities of the RNA aptamer sequences remained unknown to the experimenter until after the completion of all measurements.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	PE anti-human CD3 antibody (BioLegend, Cat #317308); Clone: OKT3; Isotype: PE Mouse IgG2a, κ Isotype Ctrl
Validation	The CD3 ϵ -positive Jurkat cell line demonstrated a robust signal upon the antibody staining, whereas the signal from the isotype control remained comparable to that of the unstained cell blank.

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	Human HEK293T (ATCC CRL-3216); Jurkat (Clone E6-1, ATCC TIB-152)
Authentication	All cell lines were authenticated by short tandem repeat (STR) profiling, with their identities also routinely monitored based on morphological features prior to use.
Mycoplasma contamination	All cell lines were tested negative for mycoplasma using a standard PCR method.
Commonly misidentified lines (See ICLAC register)	N/A

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Jurkat cells (5×10^5 cells per sample) were incubated with 5'-Cy5-labeled 2'-F-O modified RNA aptamers in 200 μ L total volume for 20 min at room temperature. Following incubation and three subsequent washes with PBS, the cells were resuspended in 100 μ L of PBS and subjected to immediate analysis using a CytoFLEX S instrument (Beckman Coulter).
Instrument	All flow cytometry data were acquired on an LSRFortessa X-20 instrument (BD Biosciences, California, USA) and a CytoFLEX S

Instrument	instrument (Beckman Coulter, California, USA).
Software	All flow cytometry data were analyzed with FlowJo V10 software (Treestar,OR,USA).
Cell population abundance	We gated the densely clustered cells for further analysis for cell purity.
Gating strategy	Instrument voltages were adjusted to optimize the distribution of cell populations on the FSC/SSC plot. Subsequent gating was performed as follows: first, cells were selected using FSC-A vs. SSC-A to exclude debris; second, single cells were discriminated by FSC-H vs. FSC-A gating; and finally, APC-positive populations were defined using fluorescence thresholds set based on unstained controls.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.