In silico evaluation of pre-training strategies based on synthetic data for functional DNA generation

Ivan Golovkin [1 2] **Stepan Druzhininskii** [1] **Daniil Litunovskiy** [1] **Amir Taldaev** [1 3 4] **Evgeny Brilenkov** [1 2] **Nikita Serov** [1]

Abstract

Despite their broad therapeutic and biotechnological potential, deoxyribozymes remain challenging to design rationally due to few consensus sequences, context-dependent activity, target inaccessibility, and a lack of large curated datasets for machine learning. This work proposes a robust multi-stage computational and experimental pipeline for DNAzyme discovery to address these limitations. The pipeline introduces domain-grounded parametric constraints for synthetic data generation, model training, and candidate evaluation. Pre-training strategies are based on two synthetic data approaches: sequences with minimal free energy boundaries matching stable structures, and sequences matching property distributions that produce structures similar to real DNAzymes. A hierarchical screening approach using machine learning models and structural priors then selects candidates for laboratory validation under conditions that include plausible cofactors crucial for catalytic activity. Overall, this pipeline provides a resource-efficient strategy for exploring novel DNAzyme catalytic cores and enabling predictable DNAzyme activity.

1. INTRODUCTION

Beyond the canonical role of genetic data storage, DNA molecules possess the capacity to function as catalysts, which was discovered by Breaker and Joyes who synthesized the first deoxyribozyme (DNAzyme) in 1994[1]. These catalytic DNA molecules are typically single-stranded sequences obtained through combinatorial *in vitro* selection techniques[2,3]. Similar to their RNA counterparts (ribozymes), both DNA and RNA can exhibit the ability to bind ligands with high affinity and specificity, and catalyze a broad spectrum of chemical reactions[4]. Initially conceived as potential gene-silencing agents due to their capacity for the sequence-specific cleavage of target mRNA, subsequent research has led to the isolation of numerous DNAzymes that facilitate a diverse array of chemical transformations. This functional diversity has significantly expanded their scope of application, firmly establishing them as genuine biocatalysts alongside ribozymes and proteinaceous enzymes[5,6].

The DNAzymes 8-17 and 10-23 are among the most well-characterized catalytic DNA molecules, renowned for their ability to mediate the sequence-specific cleavage of RNA[7]. The therapeutic applications of DNAzymes extend to a broad range of pathologies, including viral infections, cancer, and cardiovascular diseases[8–11]. Furthermore, DNAzymes represent a promising strategy for anti-inflammatory therapy. A promising avenue is anti-inflammatory therapy, exemplified in allergic asthma by the DNAzyme SB010, which targets GATA3 mRNA and has been shown to attenuate asthmatic responses in patients[12, 13]. DNAzymes also serve as valuable research tools for probing gene function, such as using a Twist-targeting DNAzyme to elucidate its role in apoptotic pathways[14–16]. Despite this promise, a central design challenge persists. While modern methods like *in vitro* selection and machine learning have revolutionized the optimization of binding arms for high-affinity, specific target recognition, the catalytic core remains a bottleneck. Designers are largely constrained to a limited set of known consensus sequences (e.g., from 8-17 or 10-23), whose efficacy is highly context-dependent and unpredictable. Optimizing this core is a laborious, empirical process that offers no guarantee of success and requires extensive resource-intensive screening. This highlights a critical need for novel, rational approaches to design *de novo* catalytic cores with high and predictable activity.

Target site inaccessibility is a common limitation for many nucleic acid-based tools, including DNAzymes. Strategies

[1]Center for AI in Chemistry, ITMO University, Saint Petersburg, Russia [2]AI Talent Hub, ITMO University, Saint Petersburg, Russia [3]Phystech school of biological and medical physics, Moscow Institute of Physics and Technology, Moscow, Russia [4]Institute of Biomedical Chemistry, Moscow, Russia. Correspondence to: Nikita Serov <nsserov@itmo.ru>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

to overcome this face significant constraints. While reaction conditions can be modified, this often lacks therapeutic relevance. Extending binding arms can impede product release, hindering rapid diagnostics. Chemically modified nucleotides enhance hybridization and stability[17–19] but require laborious optimization, costly synthesis, and are difficult to generalize for long, structured RNA (lsRNA) targets. Computational prediction of RNA accessibility loses accuracy with longer transcripts[20], and experimental mapping is often too resource-intensive for rapid development[21–23]. Using antisense oligonucleotides (ASOs) to displace obstructive structures near the cleavage site is promising but not a universal solution.

While machine learning (ML) holds transformative potential for molecular design, its application to DNAzymes remains a largely untapped frontier. Despite the introduction of foundational resources like the comprehensive DNAmoreDB[24], the field still suffers from a critical scarcity of large, well-curated, and functionally characterized datasets that are essential for training robust, predictive ML models. This data paucity is the primary bottleneck hindering the development of sophisticated computational generative tools capable of moving beyond exploratory analysis to true de novo design and reliable activity prediction.

The existing landscape of tools is notably sparse. A prominent exception is SequenceCraft[25], employs ML for the exploratory analysis of RNA-cleaving deoxyribozymes. However, while tools like SequenceCraft represent a crucial first step, the field still lacks dedicated, high-throughput platforms for the systematic generation and in silico validation of novel DNAzyme candidates. The challenge extends beyond mere sequence generation; it requires predicting complex structure-activity relationships, catalytic efficiency, and specificity within a given biological context—a task for which current datasets and algorithms are insufficient.

Generative design represents a paradigm shift in DNAzyme discovery, potentially enabling the exploration of novel catalytic centers and providing a powerful alternative to traditional SELEX procedures. However, the development of such models for functional nucleic acids remains a significant challenge. While pioneering frameworks for biomolecular design exist such as RFdiffusion[26] for protein structures and EvoDiff for generative sequence modeling, specialized[27] models for functional nucleic acids, and DNAzymes in particular, are absent. The path to developing such models is fraught with obstacles. It requires the curation of large, high-quality datasets of DNAzyme sequences with validated catalytic activity and structural information. Furthermore, model architecture must be capable of learning the complex relationship between primary sequence, higher-order structure, and catalytic function, which is poorly understood. Finally, the computational cost of training such a model and the subsequent experimental validation of its generated candidates present a substantial resource barrier that has yet to be overcome.

Therefore, there is a pressing need for two parallel advancements: first, the creation of expanded, high-quality generative datasets that capture a wider diversity of validated sequences; and second, the development of next-generation ML pipelines specifically engineered to leverage this data for the generative design of potential DNAzyme candidates. Our work proposes a robust methodology for DNAzyme discovery that comprises a multi-stage computational and experimental pipeline. The process initiates with the generation of a large-scale, general-purpose datasets for primary pre-training, followed by the selection of optimal model architectures for this task. Subsequent DNAzyme-specific fine-tuning is augmented by discriminative model filtering to select high-probability candidates. Promising sequences are further analyzed through secondary and tertiary structure prediction to identify functionally viable motifs. To infer catalytic potential, we employ clustering with known DNAzymes to hypothesize cofactor requirements, which are then evaluated via molecular docking and binding affinity simulations.

2. RESULTS AND DISCUSSION

2.1. Domain-Grounded Priors for DNAzyme Candidate Evaluation

Given the scarcity of experimentally validated DNAzyme sequences, we established qualitative and quantitative criteria to guide synthetic data generation, model training, and candidate filtering. These domain-grounded rules—analogous to Lipinski's rules in drug discovery—enable systematic evaluation of generative model outputs and provide a structured protocol for selecting high-fidelity DNAzyme candidates.

2.1.1. SEQUENCECRAFT DATASET

The SequenceCraft platform enables predictive modeling of DNAzyme activity, shifting from stochastic discovery to bioinformatic design. Its comprehensive dataset contains 349 unique catalytic cores across 549 systems, annotated with 71 reaction condition parameters. A predictive model trained on this data achieved $R^2 > 0.93$ for estimating catalytic rate constants (k_{obs}) prior to synthesis.

2.1.2. STRUCTURAL ANALYSIS

DNAzyme therapeutic function depends on three-dimensional folding that defines the catalytic core. Structural analysis identifies active sites, substrate recognition mechanisms, cofactor coordination, and motifs that enhance in vivo stability. By transitioning from primary sequences to structural models, we derived benchmark

criteria for minimal free energy (MFE), sequence length, GC content, and secondary structure elements (hairpins, loops).

MFE calculations. Catalytic efficiency is governed by thermodynamic stability. Low MFE ensures a resilient catalytic core and stable metal-ion binding at 37 °C, while high MFE indicates structural instability. In the SequenceCraft dataset, mean MFE is -7.5 ± 4.47 kcal/mol, with Mn^{2+}-associated sequences showing the lowest mean (-8.7 kcal/mol) due to Mn^{2+}'s superior coordination versatility compared to Mg^{2+}. To ensure robust catalytic conformations in our generative pipeline, we applied a conservative MFE threshold of -10.0 kcal/mol for pre-training data generation.

Secondary structure analysis. Using dot-bracket notation (parentheses for paired nucleotides, dots for unpaired regions), we characterised structural features across the dataset. Figure 1 presents distributions of sequence lengths, GC composition, hairpin center positions, and hairpin lengths.

The dataset displays a highly optimised structural profile. Sequence lengths (Fig. 1a) cluster near 40 nucleotides—sufficient for a complex catalytic core yet efficient for synthesis. GC content (Fig. 1b) centres at 45–50 %, balancing thermal stability at 37 °C against excessive rigidity. Hairpin centre positions (Fig. 1c) follow a bimodal distribution with peaks at 0.25 and 0.55, suggesting that hairpins serve either as peripheral stabilising motifs or as integral components of the catalytic core. Short hairpins (3–5 nucleotides, Fig. 1d) predominate, providing compact, rigid coordination pockets for metal ions; longer hairpins are rare, likely to avoid steric hindrance and over-stabilisation.

From these observations we derived heuristic filtering rules. Although fully unpaired sequences and stems longer than 9 nucleotides appear as rare exceptions in the reference set, their functional reliability is uncertain. We therefore excluded candidates with: (i) no paired nucleotides (unable to form a stable catalytic core), and (ii) stems > 9 nt (excessive rigidity may impair conformational flexibility during catalysis). Additionally, the normalised hairpin start and end positions exhibited a conserved distribution; we adopted the 5th and 95th percentiles (0.232 and 0.831) as selection thresholds, retaining only candidates whose hairpins fall within this window.

2.2. Synthetic Data Generation and Curation

The scarcity of experimentally validated DNAzymes limits the application of deep generative models to catalytic DNA design. Although DNA foundation models perform well in genomic annotation tasks, they are typically trained on natural genomic sequences and therefore fail to capture the biochemical and thermodynamic constraints underlying artificial catalysis. To address this limitation, we generated large-scale synthetic datasets guided by established heuristic rules and structural priors.

2.2.1. Sample preparation for model pre-training

Two independent synthetic corpora were generated to evaluate alternative strategies for modeling functional DNAzyme space. The first approach prioritized thermodynamic stability by selecting sequences with low MFE, while the second reproduced the statistical characteristics of experimentally derived datasets.

MFE based approach

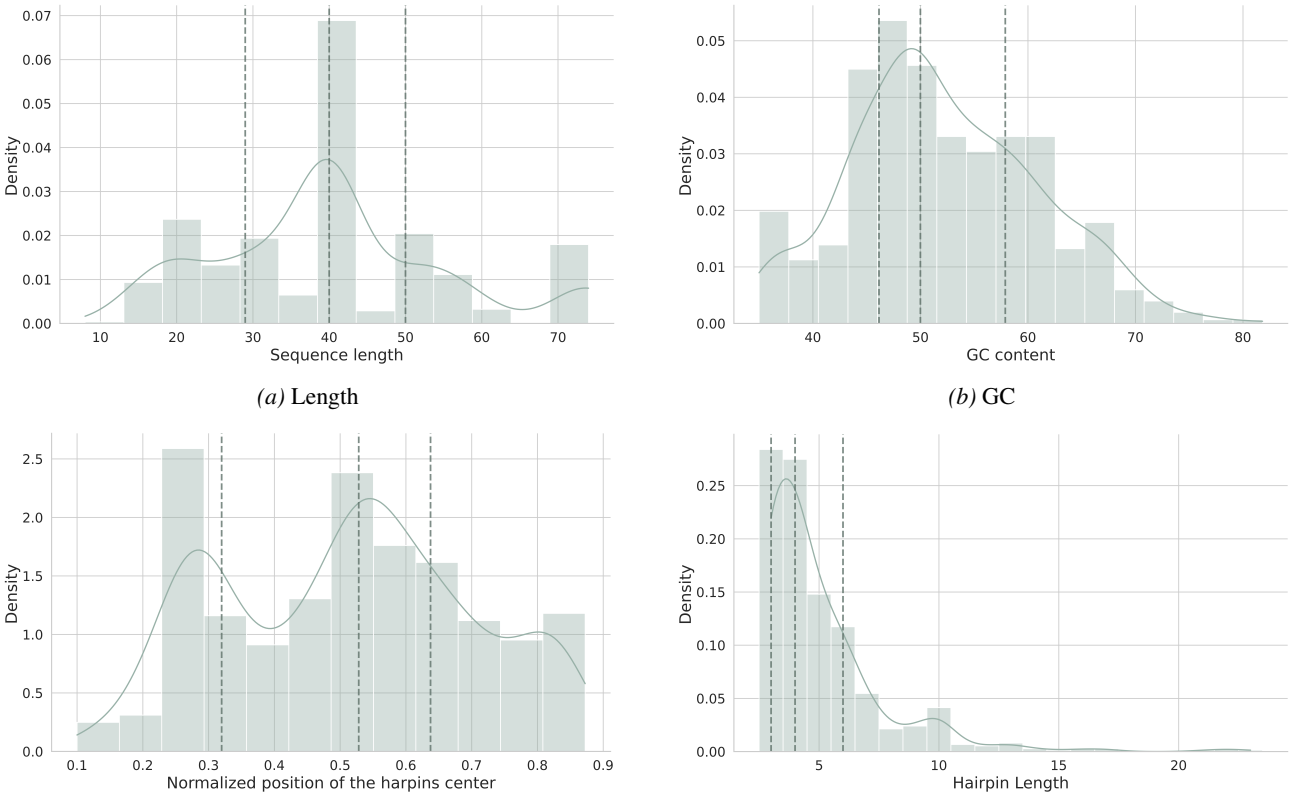
The first strategy focused on generating structurally stable sequences with MFE values below -10.0 kcal/mol. The optimization process, summarized in Algorithm 1, uses a genetic algorithm in which sequence fitness is determined by predicted secondary-structure stability. Starting from a random population, sequences with lower MFE values are preferentially retained, while poorly performing individuals are iteratively replaced through crossover and mutation operations. Selection pressure is gradually increased across generations, concentrating the search on energetically favorable regions of sequence space. The procedure continues until 250,000 sequences are generated.

Algorithm 1 MFE based approach

Initialize: Population P of random sequences, $k \leftarrow 10$, $G \leftarrow 100$, Target $N \leftarrow 250,000$
Compute Fitness: $MFE(s)$ for all $s \in P$
while Total sequences $< N$ **and** $gen < G$ **do**
  Sort P by MFE (ascending: more stable first)
  Identify P_{bottom} as the worst 90% of P
  $k \leftarrow k \times 1.1$
  $P_{parents} \leftarrow$ Select k least-fit individuals
  Apply crossover ($p = 0.75$) and mutation ($p = 0.25$)
  Recompute MFE for offspring
  Replace population and reduce its size
  $gen \leftarrow gen + 1$
end while
return stable sequences with low MFE

Distribution based approach

The second strategy aimed to reproduce the statistical profile of the SequenceCraft dataset. As outlined in Algorithm 2, reference sequences were cleaned from duplicates and outliers, after which key descriptors were extracted, including GC content, Shannon entropy, and mean Levenshtein distance. Synthetic sequences were evolved to minimize the Wasserstein distance between generated and reference fea-



(a) Length

(b) GC

(c) Hairpin position

(d) Hairpin length

Figure 1. Structural parameters of SequenceCraft dataset.

ture distributions using tournament selection, crossover, and mutation, while a hall-of-fame mechanism preserved the best candidates. This approach produced 250,000 sequences closely matching the compositional properties of the reference corpus.

The two strategies differ in computational complexity. The MFE-based method requires RNA/DNA folding with complexity $\mathcal{O}(N \cdot L^3)$, where N is the number of sequences and L is sequence length. The distribution-based method additionally computes pairwise Levenshtein distances against a reference set of size M, leading to $\mathcal{O}(N \cdot M \cdot L^2)$ complexity. For typical DNAzyme settings ($L \approx 40$, $M \approx 1000$), this makes it roughly 25× more expensive per sequence. Thus, the MFE-based strategy is more scalable, whereas the distribution-based approach requires heuristics or hardware acceleration for practical use.

Classifier

The training set consisted of 161 experimentally validated DNAzymes with RNA cleavage rates above 10^{-7} min^{-1} and 1,000 randomly generated inactive sequences. The class imbalance was intentionally introduced to minimize false positives and improve selection reliability.

Algorithm 2 Distribution based approach

Load Dataset: $D_{ref} \leftarrow$ SequenceCraft
Extract Target Distributions: $T \leftarrow \{GC, Entropy, Levenshtein\}$
Initialize: Population P with sampled sequence lengths
while Total sequences $< N$ **and** $gen < 30$ **do**
  for each sequence $s \in P$ **do**
    Compute $GC(s)$, entropy, and mean Levenshtein distance
    Evaluate Wasserstein distance to T
  end for
  Tournament selection
  Two-point crossover and shuffle-index mutation
  Preserve best-performing individuals
  $gen \leftarrow gen + 1$
end while
return Synthetic sequences matching D_{ref}

Among all tested configurations, LightGBM with DNABERT embeddings achieved the best performance. Generalization was further assessed on an independent subset of the SequenceCraft dataset excluded from training. On this benchmark, the model achieved a ROC-AUC of 0.935,

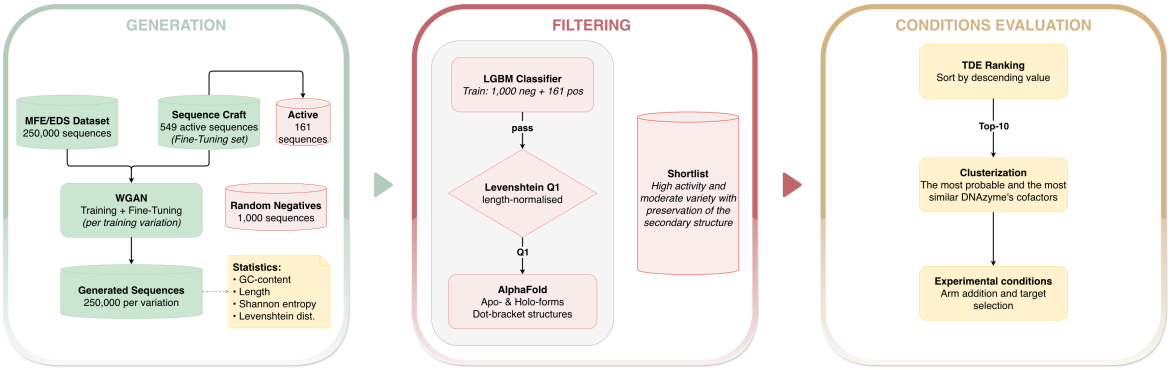


Figure 2. Hierarchical screening pipeline.

MCC of 0.788, Precision of 0.919, Recall of 0.845, and an F1-score of 0.880, with only 11 false-positive predictions. These results confirm the robustness and discriminative capability of the selected classifier.

Regressor

As an additional validation stage, generated sequences passing the classifier filter were evaluated using the regression model introduced in the SequenceCraft study. The model incorporated sequence information together with environmental parameters, including temperature, cofactor, and buffer composition. To ensure consistency, experimental conditions were fixed to the most common settings observed in the original dataset.

Clustering

To predict metal ion specificity, a k-nearest neighbors (KNN) classifier was trained on HyenaDNA embeddings of SequenceCraft sequences. The dataset was divided into 449 training and 100 test samples.

The model achieved a micro-averaged F1-score, Precision, and Recall of 0.740, indicating strong performance on dominant classes. However, macro-averaged metrics were substantially lower (F1 = 0.505, Precision = 0.523, Recall = 0.504), demonstrating reduced performance on minority classes. To compensate for this imbalance, predicted cofactors were additionally compared with those of the nearest known sequences.

Predicted metal ion labels and corresponding confidence scores were appended to each generated sequence and exported for downstream analysis.

2.2.2. Screening pipeline

A hierarchical screening pipeline was developed to reduce and prioritize generated DNAzyme candidates (Figure 2). The workflow sequentially applied: (i) classification filtering, (ii) Levenshtein distance stratification, (iii) secondary-structure filtering, and (iv) Tree Distance Editing (TDE) ranking.

The classifier served as the primary filtering stage and removed the majority of low-probability candidates. The remaining sequences were stratified by normalized Levenshtein distance to experimentally validated DNAzymes. This step balanced novelty and functional plausibility by excluding both trivial near-duplicates and highly divergent sequences. Candidates from the first quartile were retained as the closest representatives of the known functional space.

Secondary-structure filtering was then applied using rules derived from the SequenceCraft dataset. Sequences composed entirely of unpaired nucleotides were excluded due to the absence of a stable catalytic core. Candidates containing stems longer than 9 nucleotides were also removed, since excessive stabilization may reduce catalytic flexibility and increase misfolding risk. In addition, normalized hairpin coordinates were constrained to the 5th–95th percentile interval (0.232–0.831), preserving the structural organization characteristic of functional DNAzymes.

Finally, candidates were ranked using TDE. Lower TDE scores indicate secondary structures more similar to experimentally validated DNAzymes and therefore higher structural plausibility.

The multi-stage design reduces false positives associated with any single metric and ensures that selected candidates satisfy probabilistic, sequence-level, and structural constraints simultaneously.

2.2.3. The comparative analysis of approaches to training

Five WGAN training configurations were compared to evaluate the effect of synthetic pre-training on DNAzyme generation. All results of the comparison are captured in Table 3. The main objective was to determine whether pre-training on synthetic datasets generated by either MFE optimization or evolutionary distribution synthesis (EDS) improves performance relative to training only on experimentally val-

idated DNAzymes. Generated sequences from each configuration were processed through the screening pipeline, and the resulting statistics are summarized in Tables 1 and 2.

Among all configurations, EDS + FT showed the closest agreement with the validated DNAzyme distribution. It achieved the lowest Jensen–Shannon divergence (JSD) values for GC-content (0.013) and entropy (0.012), demonstrating that statistically matched synthetic pre-training effectively captures the compositional properties of functional DNAzymes.

Fine-tuning substantially improved both synthetic pre-training strategies. For MFE pre-training, fine-tuning reduced GC-content divergence from 0.063 to 0.013. In the EDS pipeline, fine-tuning reduced 3-mer JSD from 0.024 to 0.013, indicating improved local motif fidelity. Structural diversity remained stable across all models, with APJD and MPD values consistently ranging between 0.70 and 0.72 except the base MFE approach.

Negative samples (RS) were used as a control for screening performance. Surprisingly, models trained only on synthetic datasets without fine-tuning did not outperform random sequences in the number of candidates passing the screening stages. In contrast, both fine-tuned configurations produced substantially more high-quality candidates, confirming the importance of adaptation to experimentally validated DNAzymes.

Table 1. Effect of screening stages on candidate counts.

Method	EDS + FT	MFE + FT	RS	EDS	MFE	Sequence Craft
Classifier	**2191**	**1911**	600	264	343	1531
Levenshtein	**767**	**618**	151	69	99	583
Dot-bracket	**568**	**421**	108	128	46	394

The screening pipeline reduced inter-model differences by selecting structurally and compositionally conservative candidates. After filtering, compositional divergence metrics (JSD) decreased for both fine-tuned setups, whereas pairwise diversity changed only modestly: APJD remained nearly identical (0.710 for MFE + FT vs 0.709 for EDS + FT), while MPD increased slightly for both models (0.661 to 0.665 and 0.675 to 0.678, respectively). Thus, screening primarily improved compositional fidelity rather than uniformly reducing sequence diversity.

After screening, EDS + FT showed higher structural diversity (MPD 0.678 vs 0.665 for MFE + FT) and better compositional agreement with validated DNAzymes, including the lowest GC-content divergence (JSD 0.020 vs 0.038) and lower 3-mer (0.020 vs 0.029) and entropy (0.047 vs 0.091) divergence. MFE + FT, in contrast, yielded a more structurally conservative pool with poorer compositional fidelity after filtering, while also passing fewer candidates through the pipeline (421 vs 568 sequences). These results indicate that screening preserves complementary model strengths, but shifts the balance toward EDS + FT for post-filter candidate quality.

Overall, synthetic pre-training consistently improved generation quality relative to training only on scarce experimental data. Before screening, EDS + FT achieved the highest statistical fidelity to validated DNAzymes (lowest JSD values), whereas MFE + FT preserved greater structural diversity (MPD 0.725 vs 0.716). Together, these findings identify EDS + FT and MFE + FT as the most effective and complementary configurations for DNAzyme generation, with EDS + FT favoring fidelity-focused selection and MFE + FT favoring diversity-rich exploration prior to downstream filtering.

Visualizing the sequence space with Principal Component Analysis (PCA) on HyenaDNA embeddings (Figure 3) provides qualitative validation of the statistical trends described above and enables comparison between generated and natural DNAzyme distributions.

All models produced broad continuous clusters without collapse into isolated regions, indicating the absence of overfitting and successful learning of general DNAzyme patterns rather than memorization of training sequences.

The EDS + FT (Fig. 3a) and MFE + FT (Fig. 3b) configurations showed the strongest overlap with the natural DNAzyme distribution. Sequences passing the screening pipeline were concentrated within dense regions of the natural embedding space, confirming the effectiveness of the filtering strategy.

The two leading models demonstrated complementary behavior. EDS + FT showed the closest agreement with the empirical distribution, consistent with its low JSD values. In contrast, MFE + FT displayed broader dispersion in PCA space, reflecting higher structural diversity and wider exploration of thermodynamic configurations. Negative Samples formed sparse and disorganized clusters, highlighting the inefficiency of random sequence generation.

Table 2. Metrics after screening pipeline.

Setup	APJD ↑	MPD ↑	JSD ↓ (3-mer)	JSD ↓ (GC)	JSD ↓ (Entropy)
MFE + FT	0.710	0.665	0.029	0.038	0.091
EDS + FT	0.709	0.678	0.020	0.020	0.047
RS	0.718	0.717	0.007	0.034	0.015

3. Conclusion

The systematic comparison of WGAN training strategies demonstrates that pre-training on large-scale synthetic

datasets decisively outperforms training solely on the scarce collection of experimentally validated DNAzymes. Pre-trained models achieved Jensen–Shannon divergences up to an order of magnitude lower than the DNAzyme-only baseline for both global sequence properties and local motif distributions, confirming that synthetic data provide a robust statistical prior that effectively compensates for the limited availability of real catalytic sequences.

From an algorithmic perspective, the two pre-training strategies differ meaningfully in their data-generation cost. The MFE-based dataset is constructed using thermodynamic folding calculations with a time complexity of $\mathcal{O}(N \cdot L^3)$, which for short DNAzyme sequences ($L \sim 40$) is an extremely rapid operation. In contrast, the EDS pipeline requires computing Levenshtein distances to all reference sequences, scaling as $\mathcal{O}(N \cdot M \cdot L^2)$ and incurring a substantially larger computational burden as the reference set M grows. Thus, the MFE-based approach offers a compelling practical advantage: it produces a large, structurally grounded training corpus notably faster, making it the preferred option when computational efficiency and thermodynamic stability are the primary concerns.

Despite this speed benefit, the EDS-based strategy, which explicitly matches the property distributions of the Sequence-Craft dataset and incorporates subsequent fine-tuning, ultimately yielded candidates that are statistically indistinguishable from natural DNAzymes across all examined metrics, establishing EDS + FT as the method of choice for maximum biological realism. The hierarchical screening pipeline developed in this work preserved the relative quality ranking of the models: the superior fidelity of pre-trained variants persisted after filtering, with EDS + FT continuing to deliver the most realistic candidates. This confirms that the pipeline acts as a robust amplifier of inherent candidate quality rather than distorting model comparisons, while simultaneously ensuring that only the most promising sequences advance to resource-intensive experimental validation.

In summary, the synergy between computationally efficient pre-training (MFE), high-fidelity distribution-matching (EDS), targeted fine-tuning, and a principled multi-metric screening workflow constitutes a versatile, resource-efficient, and generalizable framework for the generative design of functional nucleic acids.

Accessibility

All code exits in the repository: https://github.com/GenerativeMolMachines/DNAzymesGEN

Software and Data

All data exist in the repository: https://github.com/GenerativeMolMachines/DNAzymesGEN

Author Contributions

Nikita Serov conceived the study. **Ivan Golovkin** developed the methodology, wrote the manuscript, performed visualization, and maintained the codebase. **Ivan Golovkin** and **Daniil Litunovskiy** trained the models, computed statistics, and generated datasets. **Stepan Druzhininskii** developed classifiers, validated secondary structures, and carried out clustering. **Amir Taldaev** applied AlphaFold to obtain apo and holo forms. **Evgeny Brilenkov** conducted generation experiments and adapted training scripts. **Nikita Serov** contributed to result interpretation and manuscript revision.

All authors reviewed and approved the final manuscript.

Acknowledgements

Authors thank ITMO University Research Projects in AI Initiative (RPAII) (project #640100) for infrastructure support.

References

1. Breaker, R. R. & Joyce, G. F. A DNA enzyme that cleaves RNA. *Chem. Biol.* **1**, 223–229 (1994).

2. Hollenstein, M. DNA Catalysis: The Chemical Repertoire of DNAzymes. *Molecules* **20**, 20777–20804 (2015).

3. Forty Years of In Vitro Evolution - Joyce - 2007 - Angewandte Chemie International Edition - Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1002/anie.200701369.

4. Silverman, S. K. Catalytic DNA: Scope, Applications, and Biochemistry of Deoxyribozymes. *Trends Biochem. Sci.* **41**, 595–609 (2016).

5. DNAzyme technology and cancer therapy: cleave and let die | Molecular Cancer Therapeutics | American Association for Cancer Research. https://aacrjournals.org/mct/article/7/2/243/93030/DNAzyme-technology-and-cancer-therapy-cleave-and.

6. Peracchi, A. DNA catalysis: potential, limitations, open questions. https://doi.org/10.1002/cbic.200500098 (2005) doi:10.1002/cbic.200500098.

7. Santoro, S. W. & Joyce, G. F. A general purpose RNA-cleaving DNA enzyme. *Proc. Natl. Acad. Sci.* **94**, 4262–4266 (1997).

8. DNAzymes and cardiovascular disease - Benson - 2008 - British Journal of Pharmacology - Wiley Online Library. https://bpspubs.onlinelibrary.wiley.com/doi/full/10.1038/bjp.2008.145.

9. Inhibition of bcr-abl Oncogene Expression by Novel Deoxyribozymes (DNAzymes) | Human Gene Therapy. https://www.liebertpub.com/doi/abs/10.1089/10430349950016573.

10. Goila, R. & Banerjea, A. C. Sequence specific cleavage of the HIV-1 coreceptor CCR5 gene by a hammer-head ribozyme and a DNA-enzyme: inhibition of the coreceptor function by DNA-enzyme. *FEBS Lett.* **436**, 233–238 (1998).

11. Yan, J., Ran, M., Shen, X. & Zhang, H. Therapeutic DNAzymes: From Structure Design to Clinical Applications. https://doi.org/10.1002/adma.202300374 doi:10.1002/adma.202300374.

12. Gata-3-specific Dnazyme As An Appraoch For Asthma-therapy - Journal of Allergy and Clinical Immunology. https://www.jacionline.org/article/S0091-6749(06)02366-9/fulltext.

13. Krug, N. *et al.* Allergen-Induced Asthmatic Responses Modified by a GATA3-Specific DNAzyme. *N. Engl. J. Med.* **372**, 1987–1995 (2015).

14. Caramori, G., Chung, K. F. & Barnes, P. J. Allergen Responses Modified by a GATA3 DNAzyme. *N. Engl. J. Med.* **373**, 1176–1177 (2015).

15. Grimpe, B. *et al.* The critical role of basement membrane-independent laminin gamma 1 chain during axon regeneration in the CNS. *J. Neurosci. Off. J. Soc. Neurosci.* **22**, 3144–3160 (2002).

16. Hjiantoniou, E., Iseki, S., Uney, J. B. & Phylactou, L. A. DNazyme-mediated cleavage of Twist transcripts and increase in cellular apoptosis. *Biochem. Biophys. Res. Commun.* **300**, 178–181 (2003).

17. RNA cleaving '10-23' DNAzymes with enhanced stability and activity | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/31/20/5982/1039490.

18. LNAzymes: Incorporation of LNA-Type Monomers into DNAzymes Markedly Increases RNA Cleavage | Journal of the American Chemical Society. https://pubs.acs.org/doi/full/10.1021/ja0276220.

19. Locked nucleoside analogues expand the potential of DNAzymes to cleave structured RNA targets | BMC Molecular Biology. https://link.springer.com/article/10.1186/1471-2199-7-19.

20. Bacterial Regulatory RNA: Methods and Protocols | SpringerLink. https://link.springer.com/book/10.1007/978-1-61779-949-5.

21. Rapid in vitro Method for Obtaining RNA Accessibility Patterns for Complementary DNA Probes: Correlation with an Intracellular Pattern and Known RNA Structures | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/25/24/5010/1748765.

22. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/42/21/e165/2903199.

23. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution | Nature Protocols. https://www.nature.com/articles/nprot.2006.249.

24. Ponce-Salvatierra, A., Boccaletto, P. & Bujnicki, J. M. DNAmoreDB, a database of DNAzymes. *Nucleic Acids Res.* **49**, D76–D81 (2021).

25. Eremeyeva, M., Din, Y., Shirokii, N. & Serov, N. SequenceCraft: machine learning-based resource for exploratory analysis of RNA-cleaving deoxyribozymes. *BMC Bioinformatics* **26**, 2 (2025).

26. Watson, J. L. *et al.* Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. 2022.12.09.519842 Preprint at https://doi.org/10.1101/2022.12.09.519842 (2022).

27. Protein generation with evolutionary diffusion: sequence is all you need | bioRxiv. https://www.biorxiv.org/content/10.1101/2023.09.11.556673v1.

28. Breaker, R. R. In vitro selection of catalytic polynucleotides. *Chem. Rev.* **97**, 371–390 (1997).

29. Silverman, S. K. Deoxyribozymes: selection design and serendipity in the development of DNA catalysts. *Acc. Chem. Res.* **42**, 1521–1531 (2009).

30. Zhao, Z. *et al.* Structural basis for catalysis by a DNAzyme. *Nat. Chem.* **8**, 684–690 (2016).

31. Chen, J. *et al.* Single-molecule dynamics of a DNAzyme reveals hidden states. *Nat. Commun.* **8**, 1–11 (2017).

32. Herschlag, D. RNA chaperones and the RNA folding problem. *J. Biol. Chem.* **270**, 20871–20874 (1995).

33. Jencks, W. P. *Catalysis in Chemistry and Enzymology*. (Dover Publications, 1987).

34. Fedor, M. J. & Williamson, J. R. The catalytic diversity of RNAs. *Nat. Rev. Mol. Cell Biol.* **6**, 399–412 (2005).

35. Liu, J. & Lu, Y. Rational design of functional DNA-based metal sensors. *Angew. Chem. Int. Ed.* **45**, 90–94 (2006).

36. Li, J. & Breaker, R. R. Kinetics of RNA degradation by specific DNA enzymes. *J. Am. Chem. Soc.* **121**, 5364–5372 (1999).

37. Peracchi, A. Origins of the temperature dependence of ribozyme catalysis. *Biochemistry* **44**, 10542–10552 (2005).

38. Takahashi, S., Hamada, M., Tateishi-Karimata, H. & Sugimoto, N. Fitness landscapes and thermodynamic approaches to development of nucleic acids enzymes: from classical methods to AI integration. *RSC Chem. Biol.* **6**, 1667–1685 (2025).

39. Zhang, F., Shi, W., Guo, L., Liu, S. & He, J. The Programmable Catalytic Core of 8-17 DNAzymes. *Molecules* **29**, 2420 (2024).

40. Hollenstein, M., Hipolito, C., Lam, C., Dietrich, D. & Perrin, D. M. A self-cleaving DNA enzyme modified with amines, guanidines and imidazoles operates independently of divalent metal cations (M2+). *Nucleic Acids Res.* **37**, 1638–1649 (2009).

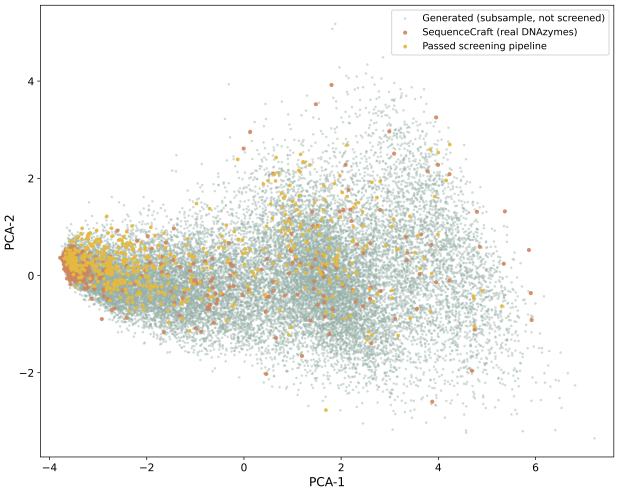
A. Apendix

B. Additional tables

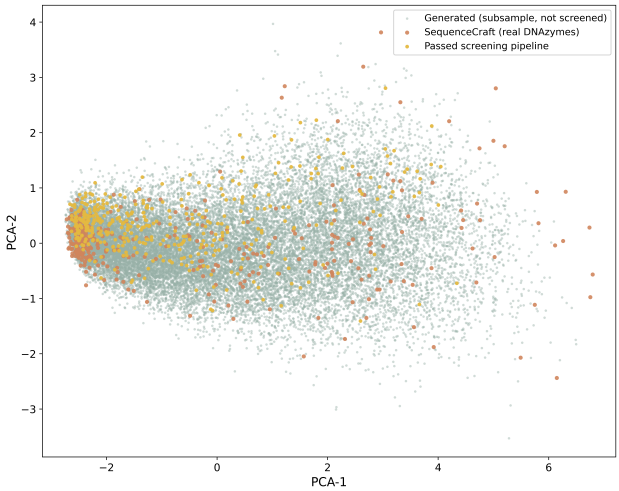
Table 3. Performance metrics of WGAN architectures.

Training setup	APJD ↑	MPD ↑	JSD ↓ (3-mer)	JSD ↓ (GC)	JSD ↓ (Entropy)
Negative	0.717	0.709	0.013	0.035	0.023
SequenceCraft	0.713	0.708	0.003	0.014	0.007
MFE	0.649	0.513	0.013	0.063	0.062
MFE + FT	0.725	0.725	0.007	0.013	0.016
EDS	0.714	0.691	0.015	0.024	0.025
EDS + FT	0.718	0.716	0.004	0.013	0.012

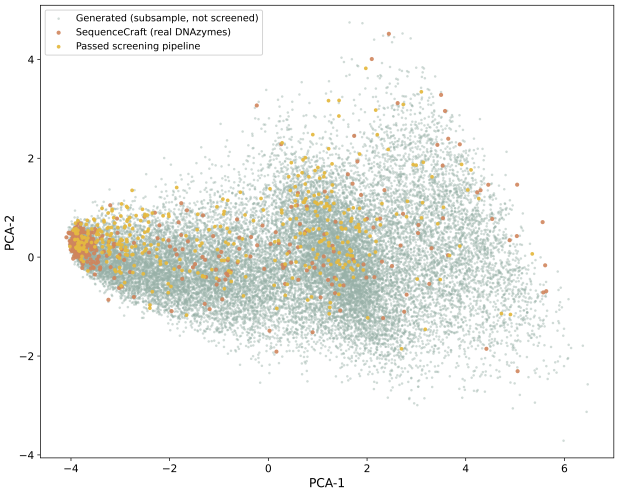
C. Additional figures



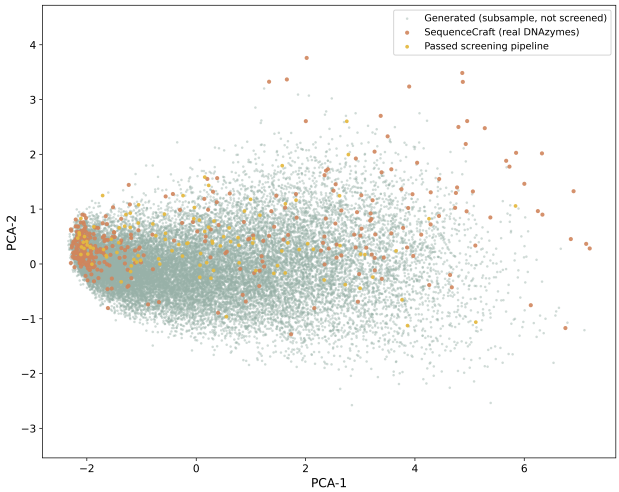
(a) EDS + FT



(b) MFE + FT



(c) DNAzymes



(d) Negative Samples

Figure 3. PCA projection of generated and reference DNAzyme sequences using HyenaDNA embeddings.