BMC Bioinformatics

SOFTWARE Open Access

# SequenceCraft: machine learning-based resource for exploratory analysis of RNA-cleaving deoxyribozymes

M. Eremeyeva[1], Y. Din[1], N. Shirokii[1] and N. Serov[1*]

*Correspondence:
serov@scamt-itmo.ru

[1] International Institute "Solution Chemistry of Advanced Materials and Technologies", ITMO University, Saint-Petersburg, Russian Federation 191002

## Abstract

**Background:** Deoxyribozymes or DNAzymes represent artificial short DNA sequences bearing many catalytic properties. In particular, DNAzymes able to cleave RNA sequences have a huge potential in gene therapy and sequence-specific analytic detection of disease markers. This activity is provided by catalytic cores able to perform site-specific hydrolysis of the phosphodiester bond of an RNA substrate. However, the vast majority of existing DNAzyme catalytic cores have low efficacy in in vivo experiments, whereas SELEX based on in vitro screening offers long and expensive selection cycle with the average success rate of ~ 30%, moreover not allowing the direct selection of chemically modified DNAzymes, which were previously shown to demonstrate higher activity in vivo. Therefore, there is a huge need in in silico approach for exploratory analysis of RNA-cleaving DNAzyme cores to drastically ease the discovery of novel catalytic cores with superior activities.

**Results:** In this work, we develop a machine learning based open-source platform SequenceCraft allowing experimental scientists to perform DNAzyme exploratory analysis *via* quantitative observed rate constant ($k_{obs}$) estimation as well as statistical and clustering data analysis. This became possible with the development of a unique curated database of > 350 RNA-cleaving catalytic cores, property-based sequence representations allowing to work with both conventional and chemically modified nucleotides, and optimized $k_{obs}$ predicting algorithm achieving $Q^2 > 0.9$ on experimental data published to date.

**Conclusions:** This work represents a significant advancement in DNAzyme research, providing a tool for more efficient discovery of RNA-cleaving DNAzymes. The SequenceCraft platform offers an in silico alternative to traditional experimental approaches, potentially accelerating the development of DNAzymes.

**Keywords:** Deoxyribozymes, Machine learning, Catalytic activity, Artificial intelligence, Web resource

## Background

Deoxyribozymes, also known as DNAzymes, are short single-stranded DNA molecules that exhibit catalytic activity across a wide range of chemical reactions. A particularly promising application of DNAzymes lies in their capacity for gene silencing *via* specific binding and inhibition of target RNA [1]. This attribute positions DNAzymes as valuable tools in nucleic acid-based therapeutic modalities along with ribozymes, antisense oligonucleotides (ASO), and small interfering RNA (siRNA). Moreover, their RNA-cleaving ability renders them effective in biosensing applications for the detection of metal ions, bacteria, and viruses [2–4]. In comparison to RNA- and protein-based catalysts, they offer distinct advantages e.g. small molecular size, exceptional stability, and cost-effectiveness [5]. Notably, research has demonstrated the efficacy of DNAzymes in modulating the expression of target genes implicated in various pathological conditions, including cancer, cardiovascular diseases, bacterial and viral infections, as well as central nervous system [6].

Discovery of novel catalytic nucleic acids is achieved through the Systematic Evolution of Ligands by EXponential enrichment (SELEX) methodology screening for catalytically active sequences among approximately $10^{15}$–$10^{20}$ random sequences [7]. While SELEX offers a straightforward approach, it is labor-intensive and costly due to the need for multiple selection cycles. Furthermore, the exponential increase in potential candidates with longer randomized regions poses a challenge in exploring and identifying active DNAzymes. Notably, the success of SELEX in generating functional DNAzymes is not guaranteed in every instance and estimated to be ~ 30%, underscoring the importance of thorough screening and optimization in the selection process [8].

In addition to SELEX, alternative approaches have been explored for the discovery of active DNA catalysts. Research has shown that DNAzymes can be developed through test tube evolution from a specific DNA sequence [9]. In this methodology, a non-catalytic single-stranded DNA sequence is selected as the starting point, as opposed to a random-sequence DNA pool. Throughout each selection cycle, slight mutations are introduced into the sequence *via* DNA amplification, leading to the gaining of catalytic activity by some of these mutated DNA molecules.

With their catalytic activity depending drastically on single nucleotide substitutions, chemistry of cofactor, experimental conditions, and other factors, DNAzymes represent a complex multiparametric system hard to develop and optimize *via* experiment. Thus, development of specialized computational tools is necessary to simplify the DNAzymes research and expand their applications in practice. In recent years, significant progress has been made in the development of the bioinformatics approach to DNAzymes. Ponce-Salvaterra et al. have compiled a database of published DNAzyme-related data known as DNAmoreDB [10], which aggregates information regarding sequence composition, selection conditions, catalyzed reactions, kinetics, substrates, cofactors etc. and serves as a comprehensive resource for researchers in this field. This database provides a structured view of over 1,700 DNAzymes with various catalytic activities, facilitating access to a wealth of published information within a single repository. However, the lack of negative examples of sequences without catalytic activities, especially lost its activity due to nucleotide substitutions and mutations, limits the use of data-driven approaches for novel DNAzymes discovery.

The same research group has developed a web application DNAzymeBuilder for the selection of the most effective RNA- and DNA-cleaving DNAzymes for a user-defined substrate [11]. The application utilizes a search algorithm that identifies k-mers, short specific sequences of nucleotides, which match a predetermined recognition site within its database. Manual design of DNAzymes is a time-intensive process and open to human error, therefore, an automated tool for constructing optimal ones can greatly help researchers in this field. However, it is important to note that the algorithm is constrained to searching within the known DNAzyme space and is limited to the examples stored in the database.

In addition to the DNAzymes repository and DNAzymes selector detailed above, an energy-based method has been developed to estimate catalytic activity based on heteroduplex stability of various 10–23 DNAzyme binding arms [12]. A logistic regression model was constructed using data from a limited sample size of 15, thus limiting its widespread applicability. This model was successful in identifying efficient HPV16 E6/E7-targeting DNAzymes based on parameters such as $\Delta G$, hairpin energy, and dimer energy. However, an efficient DNAzyme targeting HPV36 was erroneously predicted to be inactive due to its low dimer energy.

In nucleic acid research, there has been a recent shift towards utilizing computational methods, although only preliminary steps have been taken in the development of a tool for in silico selection of novel DNAzymes, considering all pertinent activity-influenced parameters. The selection of binding arms is largely based on complementarity to the selected substrate, and varying only the arms leads to the fact that the resulting DNAzymes effectiveness is limited by existing catalytic cores.

The central component of a DNAzyme is its catalytic core, which currently needs discovery through either an in vitro selection process or exploration of single-nucleotide mutations [13]. This study proposes a comprehensive in silico screening strategy for RNA-cleaving DNAzyme catalytic cores, based on machine learning (ML) algorithms capable of predicting DNA sequence rate constants based on various sequence-, cofactor-, and buffer-related factors. In this paper we introduce SequenceCraft, an extensible open-access platform comprising a curated database of RNA-cleaving DNAzymes, catalytic activity-predicting algorithm, and visualization tools, facilitating the preliminary in silico assessment of potential DNAzyme candidates' activity. The platform is accessible at https://sequencecraft.aicidlab.itmo.ru/.

## Implementation

### Data collection

The primary data source utilized in this study is the Application Programming Interface (API) of the DNAmoreDB [10], a database dedicated to DNAzymes. Data extraction was focused on DNAzymes capable of RNA cleavage, with the retrieved information being saved in CSV format. A similar approach was adopted for acquiring data on RNA-ligating DNAzymes. Each dataset entry encompassed details such as the DNAzyme name, sequence as reported in the respective literature, substrate, reaction conditions, kinetic properties, and comprehensive source publication information. To consider the dependence of the rate constant on temperature, we also manually collected experimental temperature values for those DNAzymes for which the kinetic characteristics are known.

Integration of Digital Object Identifiers (DOIs) from the Crossref Commons package facilitated the inclusion of direct links to the original publications for user reference. Subsequent data analysis was conducted using the dataset obtained on this step.

Following the extraction of RNA-cleaving DNAzymes, a total of 1,028 samples were obtained, out of which only 178 featured kinetic data. Among these, 16 DNAzymes had graphical characterization in the articles of the $k_{obs}$ under varying experimental conditions, including cofactor type and concentration, pH, and temperature. Data points on the plot were semi-automatically extracted using WebPlotDigitizer desktop software [14], resulting in the addition of 207 rows to the dataset where all values except the parameter changed on the graph were similar to the original sample.

#### Data preprocessing

The target variable, which is $k_{obs}$, was initially presented in a string format. Using regular expressions, the numerical values were extracted and standardized to one measurement unit. Additionally, string values detailing the composition of the buffer were disaggregated into individual components and converted to moles per liter using regular expressions. Subsequently, the concentration of each buffer component was treated as a distinct parameter. Components present in less than 20% of the samples were excluded from the analysis. For the remaining components, missing values were imputed with zeros to indicate their absence in the buffer solution. Furthermore, metal ion or compound concentrations were merged into a one parameter termed "cofactor concentration". To characterize the properties of the cofactor, four parameters of metal ions (electron affinity, ionic radii, first ionization energy of the element, nuclear charge) sourced from the Python Materials Genomics (pymatgen) library [15] were employed. In cases where multiple cofactors were utilized, an average of these parameters among the presented cofactors was calculated. Conversely, if no cofactor was required, both cofactor parameters and concentration were set to 0.

#### Data analysis

The diversity of unique DNAzymes present in the dataset was investigated by calculating Levenshtein distances between all sequences. Subsequent visualization was conducted using t-distributed stochastic neighbor embedding (t-SNE), a statistical technique for visualizing high-dimensional data in two or three dimensions. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was then applied to distribute the new two-dimensional data into clusters. Additionally, multiple sequence alignments were carried out using the web-based version of the CLUSTALW software.

#### Feature extraction

In preparation for subsequent analysis and model training, only DNAzymes with known values of the observed reaction rate constant ($k_{obs}$) were included in the dataset.

Five distinct approaches to DNA representation were compared:

(1) For one-hot representation DNA sequences were encoded using a scheme, where each nucleotide was mapped as follows:

A > [1, 0, 0, 0]
C > [0, 1, 0, 0]
G > [0, 0, 1, 0]
T > [0, 0, 0, 1]

Any non-(A, T, C, G) characters were encoded as [1, 1, 1, 1]. To address variations in sequence lengths, tensors were padded with [0, 0, 0, 0] values at the end of sequence to match the maximum observed sequence length.

(2) Features such as 2,3- and 4-mers vectors were computed using the PyBioMed [16] package and PyDNA [17] modules.

(3) The calculation of secondary structure descriptors was performed using the Python seqfold package, which leverages mfold for predicting RNA secondary structures. The dot-bracket representation generated by this library was encoded as follows:

( > −1
. > 0
) > 1

(4) A pre-trained large language model named HyenaDNA with 1 k parameters (tiny-1 k) was leveraged to generate contextual DNA embeddings [18].

(5) Furthermore, a novel convolutional autoencoder (CAE)-based representation approach was introduced.

#### Feature selection

In the process of selecting optimal parameters for system characterization, a thorough analysis was conducted to compare various descriptors and their potential combinations for the given sequence. Specifically, for HyenaDNA embeddings, an investigation was carried out to assess the feasibility of compressing the data into a 94-dimensional vector while maintaining 95% of the explained variance through the application of Principal Component Analysis (PCA). Two models were selected for evaluation, feature selection and subsequent experimentation involving the inclusion of inactive sequences and descriptors describing the buffer. The chosen models were Random Forest (RF) and Light Gradient Boosting Machine (LightGBM). The model training pipeline was systematically structured to include the concatenation of descriptor vectors, the allocation of a test sample comprising 20% of the entire dataset stratified by $k_{obs}$ value, hyperparameter optimization via a random search algorithm within a threefold stratified cross-validation (CV) framework, and further validation through a fivefold stratified CV process to compute the $Q^2$ metric. Subsequently, the $R^2$ and RMSE metrics were calculated based on the test sample.

### Model training and validation

Upon completion of the dataset formation process encompassing a comprehensive array of descriptors and data, a ML model screening process was conducted to identify the optimal model. Three gradient boosting models, namely XGBoost, LightGBM, and CatBoost, in addition to a RF model, were employed for this purpose. Prior to model evaluation, each algorithm underwent an hyperparameter optimization procedure using Randomized Search CV. Subsequently, a validation dataset comprising 53 unique sequences, distinct from the training (396 samples) and test (100 samples) sets, was curated. Feature scaling was executed through the application of the Standard Scaler. The quality of the models was evaluated using the $R^2$ and RMSE on test and validation sets as well as $Q^2$ metric through a 5-fold stratified cross-validation approach on the training set. The best LightGBM model underwent a feature selection process *via* Recursive Feature Elimination, resulting in the identification of 24 key features considered critical for model performance. The importance of these selected features was meticulously assessed using the feature_importances_ attribute, which computes the mean and standard deviation of impurity to decrease accumulation within each tree. The final model's performance was evaluated across the validation, training, and test sets.

### Web resource

For the sake of ML models integration, the application is built with Python 3 programming language on the stack of Django framework and PostgreSQL database, using libraries such as biopython and forgi for sequence-related analysis.

## Results and discussion

### Data collection and analysis

The data collection phase represents a critical stage in the development of predictive models, as the model capacity to address real-world issues faced by experimental scientists and its ability to generalize to novel instances are determined by the quality and diversity of the data obtained. Insufficient quantities of high-quality data can result in diminished accuracy of the model when applied to practical scenarios.

During the research, several sources of DNAzyme-related data were used, where the primary source was the recently released DNAzyme repository, known as DNAmoreDB [10]. In the database, three parameters characterizing DNAzyme catalytic activity are presented, namely reaction yield (71 samples), turnover number ($k_{cat}$; 14 samples), and observed reaction constant ($k_{obs}$; 154 samples). Among these values, the $k_{obs}$ emerges as the most suitable choice for direct prediction based on both data availability and physical interpretation standpoint (complete list of the extracted parameters used later on can be found in Table S1). It is important to note that the data collected to date in this field is sufficient for exploratory analysis of DNAzyme catalytic cores, however, not containing parameters crucial for fully quantitative prediction e.g., sufficient amount of single nucleotide polymorphism (SNP) studies as well as catalytic activity dependence on dinucleotide cleavage site composition, showing only the maximal activity achieved with the particular catalytic core. Since the paper focuses on the determination of intrinsic DNAzyme catalytic core ability to perform the cleavage reaction and discovery of potentially active candidates, substrate-binding arms as well as RNA substrates were

not used. High bias towards the most frequent RNA substrates observed in published structured data will render determination of catalytic ability of DNAzyme core itself difficult. Thus, experimental binding arms optimization given specific RNA substrate is still needed for discovered candidates. Findings presented further are related to the prediction of maximal catalytic activity achievable for a given group of related DNAzyme catalytic cores, where the development of SNP level precise algorithms requires more experimental mutation studies.

A collection of 1,028 DNAzymes (raw dataset prior to preprocessing) sourced from 56 scientific articles was extracted comprising DNAzymes with RNA-cleaving activity. Since DNAmoreDB serves only to aggregate existing DNAzyme-related data for experimental scientists, a significant portion of these DNAzymes lacks quantitative activity data, where 154 DNA sequences with known $k_{obs}$ were meticulously selected for further analysis. Each sequence underwent precise validation against the original source, with adjustments made to the catalytic core sequence to exclude binding arms. Additionally, buffer conditions were fine-tuned to correspond to the counter-measured value of $k_{obs}$ and were supplemented with experimental temperature data. Notably, for 16 of the DNAzymes graphical representations depicting the relationship between the $k_{obs}$ and various buffer parameters such as temperature and pH were available in the primary literature. Subsequently, this data was converted to numeric and encoded, resulting in an additional 207 samples (total 361 labeled samples) which were instrumental in enhancing the performance of the model on the final steps.

It is important to perform a comprehensive data analysis of the final database to determine limitations of the future models related to parameter ranges present, as well as to extract meaningful dependencies of the target value on experimental parameters. To better understand the range of the constants under consideration, it is crucial to assess its distribution as it directly impacts the ability of ML models to predict $k_{obs}$ in a wide range for unknown sequences. It can be seen on Fig. 1a that the distribution of $k_{obs}$ is close to normal and centered around 0.01 $min^{-1}$, with minimal and maximal values equal to 0.0001 and 1.7 $min^{-1}$, respectively. Therefore, collected dataset covers a wide range of $k_{obs}$ values, from almost inactive to very active DNAzyme catalytic cores, proving dataset completeness in a sense of $k_{obs}$ and future ML models' ability to work with DNAzymes of varying potential catalytic activity.

To assess the generalizability of future ML models, not only $k_{obs}$ coverage analysis but also a comprehensive analysis of the sequence's diversity is crucial to prove dataset completeness in a sense of sequential information. To estimate sequence diversity within the dataset, pairwise Levenshtein distances were computed quantifying the dissimilarity between two sequences by measuring the minimum number of single-nucleotide edits (insertions, deletions, or substitutions) required to transform one sequence to another. A $154 \times 154$ distance matrix was generated, compressed using t-SNE, and analyzed using scatter plots (Fig. 1b), resulting in the identification of three distinct clusters. Notably, sequence length emerged as a key differentiating parameter among the clusters, with median values of 40, 52, and 26 for clusters 0, 1, and 2, respectively (Table S2). No clear separation of clusters by $lg(k_{obs})$ was observed (Fig. S1), where a notable variance of $lg(k_{obs})$ values within each cluster (from 0.79 to 1.15) was observed. Although sequences within the same cluster exhibited relative positional proximity, multiple sequence

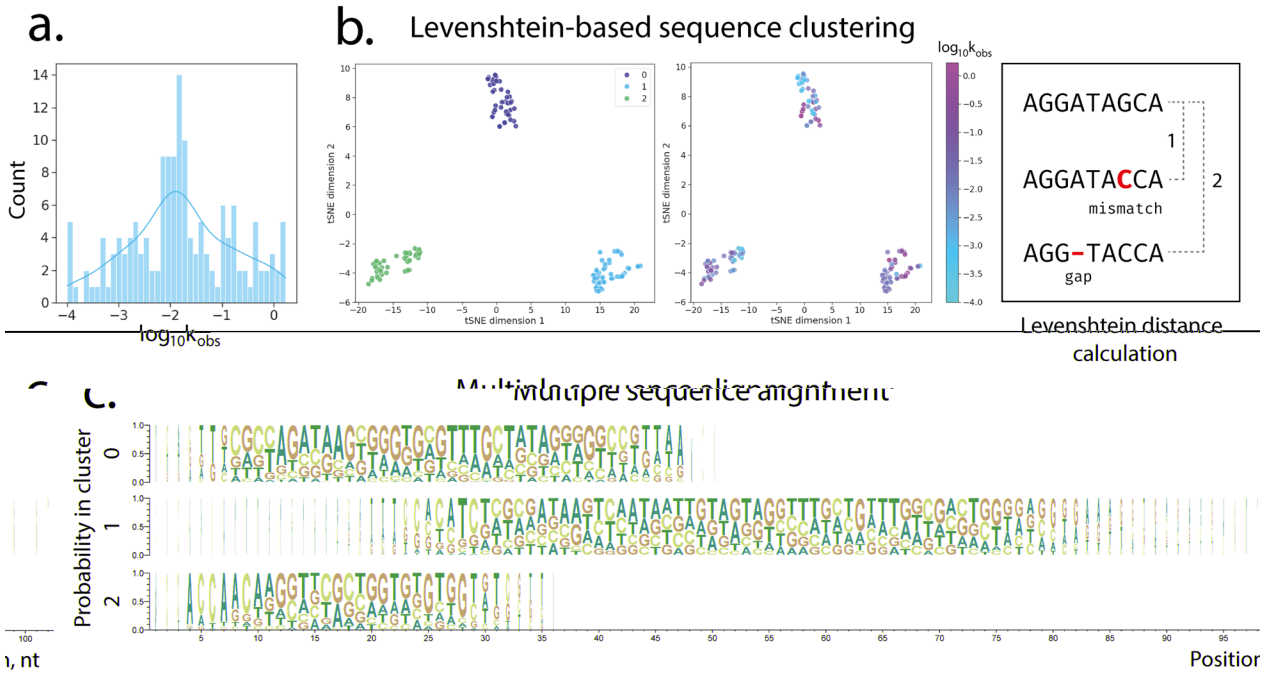


**Fig. 1** Data analysis: (**a**) $k_{obs}$ distribution (in logarithmic scale), (**b**) t-SNE clustering on pairwise Levenshtein distances (colored by cluster membership and lg($k_{obs}$) values, respectively), (**c**) consensus sequences after multiple sequence alignment for clusters 0 (violet), 1 (light blue), and 2 (green) separately formed during t-SNE clustering. Inset —Levenshtein distances calculation example

alignment revealed no dominant pattern for each cluster except for cluster 2 with high GC content (Fig. 1c). Notably, no consensus sequence was discernible; instead, localized regions displayed a prevalence of specific nucleotides.

This observation suggests that the dataset comprises diverse DNAzyme catalytic core sequences, spanning lengths ranging from 8 to 74 nucleotides. Therefore, from the analysis it can be concluded that the collected dataset contains highly diverse DNAzyme catalytic core sequences. At the same time, the absence of clear separation of sequences by $k_{obs}$ suggests that, despite its importance for catalytic activity, only sequential data will be not enough to estimate $k_{obs}$, as even closely related sequences exhibit significant variance in $k_{obs}$. This observation aligns with previous studies involving mutational analysis of certain DNAzymes [19], wherein critical substitutions in the catalytic center were identified as essential for the activity and capable of inducing critical impairment.

#### DNA features comparison

Since even subtle modifications to DNAzyme catalytic cores change $k_{obs}$ drastically, it is crucial to choose DNAzyme sequence descriptors able to consider positional information to be able to predict $k_{obs}$ in a quantitative manner. Existing DNA sequence descriptors can be categorized into two main groups: (1) bioinformatics-based, quantifying positional information for sequences utilizing statistical approaches (k-mers and one-hot encoding, which demonstrated efficacy in various genomic sequence classification tasks) [20]; (2) Large Language Model (LLM)-based, forming latent representations of sequential data (HyenaDNA model outperforming analogous LLMs across multiple benchmark assessments) 18. In addition to existing sequence representations, we implemented a dot-bracket notation of secondary structures calculated using mfold which was converted to a numerical vector [21]. The last approach, namely sequential compression of nucleotide physicochemical properties with autoencoder (AE) architecture, was implemented in this paper for the first time. Briefly, DNA sequences were transformed into matrices where each column represented properties of a specific nucleotide at a given position in the sequence. Thus, such model architecture allows to process not only conventional but also chemically modified nucleotides, which allows to use this approach on a wider range of nucleic acid structures. The convolutional autoencoder (CAE) model was trained on a dataset containing over 500,000 unique DNA sequences having non-zero Levenshtein distance to each other. The latent space vectors obtained from this model were utilized as sequence representations. Finally, it was checked on the ability to differentiate between proteins, RNA, and DNA indirectly from these latent space-derived descriptors (Fig. S2) showing great learning performance.

Each of the representation methods mentioned above offers unique advantages and disadvantages that are tailored to specific problems. Thus, to determine the most suitable descriptors for DNAzyme sequence representation, all mentioned descriptors were evaluated in sequence-only $k_{obs}$ prediction task using tree ensemble boosting models recently proven to outperform other models on tabular data, which are Random Forest (RF) and Light Gradient Boosting Machine (LightGBM) [22]. A total of 154 samples with known $k_{obs}$ were utilized, with 20% allocated to the test set and 80% to the training set. Stratified $k_{obs}$ sampling was employed to ensure a balanced distribution of the target variable across both sets. Subsequently, two models, RF and LightGBM, with pre-optimized

hyperparameters, were trained on the training set. The performance of these models was evaluated through tenfold CV and test set metrics, and the average of two models' performance was used as descriptor performance assessment metrics.

Despite exhibiting strong performance in CV (Table 1), bioinformatics-based descriptors demonstrated a substantial decline in their ability to characterize an independent test set, with performance decreasing by more than half for all features except 2-mers ($Q^2 = 0.181$, $R^2$ test $= 0.146$). Notably, 2-mers emerged as the most effective representation within the group of bioinformatics approaches. In contrast, HyenaDNA embeddings, which were compressed with principal component analysis (PCA) before use, surpassed the aforementioned features when employing the LLM-based approach, as evidenced by better performance in CV as well as on a test set ($Q^2 = 0.207$, $R^2$ test $= 0.179$). This improvement may be attributed to the model enhanced capability to capture contextual information from the sequence facilitated by the attention mechanism. Of particular significance is the performance of proposed autoencoder, which emerged as the top-performing feature ($Q^2 = 0.26$, $R^2_{\text{test}} = 0.27$), demonstrating consistent efficacy in both CV and on the test set. This outcome underscores the importance of incorporating physicochemical properties of nucleotides into the mathematical representation of oligonucleotides. Although we anticipated that the secondary structure of the catalytic core would significantly improve predictive accuracy, descriptors based on dot-bracket notation only performed poorly ($Q^2 = -0{,}072$, $R^2$ test $= 0.073$). However, according to thermodynamic hypothesis or Anfinsen's dogma primary structure fully determines the spatial folding of biological polymers to secondary and tertiary structures; therefore, the use of primary structures descriptors is sufficient for indirect yet proper spatial structure consideration by the model.

Despite top-performing DNAzyme catalytic core sequence descriptors in the task of sequence-only $k_{obs}$ prediction were established with the dominance of property-based descriptors, future ML model ability to perform exploratory analysis as well as predict $k_{obs}$ with single nucleotide resolution can be improved using several types of descriptors covering several different aspects of DNAzyme catalytic core sequences. Thus, subsequent analysis involved the comparison of combinations of the most effective features from each group (Table 2), as they contain distinct information that can complement each other, thereby enhancing the overall information content. Each combination exhibited an increase in the coefficient of determination during CV and on the test set.

**Table 1** Comparison of DNAzyme catalytic core sequence descriptors in sequence-only $k_{obs}$ prediction task

| Type | Descriptors | $Q^2$ | $R^2_{test}$ | $RMSE_{test}$ |
|---|---|---|---|---|
| Bioinformatics-based | 2-mers | 0.181 | 0.146 | 0.942 |
| | 3-mers | 0.214 | 0.076 | 0.980 |
| | 4-mers | 0.169 | 0.058 | 0.985 |
| | one-hot | 0.150 | −0.138 | 1.082 |
| Secondary structure-based | encoded dot-bracket | −0.072 | 0.073 | 0.969 |
| LLM-based | HyenaDNA + PCA | 0.207 | 0.179 | 0.923 |
| Property-based | AE embeddings | 0.259 | 0.267 | 0.871 |

**Table 2** Comparison of top-performing descriptor combinations in sequence-only $k_{obs}$ prediction task

| Descriptors combination | $Q^2$ | $R^2_{test}$ | $RMSE_{test}$ |
|---|---|---|---|
| HyenaDNA PCA<br>AE | 0.275 | 0.268 | 0.872 |
| AE<br>2-mers | 0.334 | 0.335 | 0.829 |
| HyenaDNA PCA<br>2-mers | 0.233 | 0.207 | 0.908 |
| 2-mers<br>AE<br>HyenaDNA PCA | 0.318 | 0.298 | 0.853 |

Notably, the amalgamation of all three descriptors did not yield superior results, being outperformed by the AE-2-mers combination ($Q^2 = 0.33$, $R^2_{test} = 0.33$). The inclusion of a vector of k-mers resulted in a notable 26% enhancement in model performance when compared to solely utilizing CAE descriptors. Therefore, the combination of the proposed AE model and 2-mers has shown the highest performance in sequence-only $k_{obs}$ prediction task.

Following the analysis of various descriptors, it is important to note that the specificity of the selected features lies in their aggregating nature; specifically, the AE descriptors generalize information through convolution as well as k-mers. However, this method has a notable limitation: it does not reliably discriminate between single-nucleotide mutations. For instance, when predicting the activity of single-nucleotide mutations in the catalytic core 10–23, the predicted values exhibited a narrow spread of only 0.003 $min^{-1}$ (Fig. S3a), despite some mutations leading to a complete loss of activity in experiment. This indicates that while the descriptors suggest these sequences have varying activities, they fail to accurately represent the scale of these differences and at present can be used for semi-quantitative exploratory analysis only. It is important to highlight also that there is a lack of ability to cluster samples with single-nucleotide mutations into active and inactive categories based on the selected sequence descriptors (Fig. S3b). A contributing factor to this issue is that many single-nucleotide mutations for most catalytic cores are underrepresented in the literature being presented only for several most frequently used catalytic cores, which places greater emphasis on sequence diversity. Consequently, while this approach enhances the model's generalization capabilities, it compromises its ability to recognize point mutations effectively.

#### Handling data imbalance problem

One significant challenge associated with collected data is the absence of negative examples, which can lead to overestimation of $k_{obs}$ for less active DNAzyme catalytic cores. To address this imbalance in distribution, training datasets must include sequences that are inactive in order to presume $k_{obs}$ of less active and inactive sequences more correctly. Here we explore two methodologies for simulating inactive sequences and attributing them an RNA cleavage rate constant of $10^{-7}$, a value indicative of the rate of spontaneous RNA cleavage under physiological conditions [23].

To prevent the inclusion of inactive sequences due to their unusual lengths, we generate random DNA sequences with lengths similar to those of RNA-cleaving DNAzymes. However, from a modeling perspective, this approach has proven to be less effective, as the model fails to discern between random sequences and DNAzymes. The introduction of inactive samples into the dataset leads to a significant decline in model performance, followed by erratic fluctuations (Fig. 2). This behavior may stem from the model's tendency to overfit when presented with predominantly inactive sequences. The second approach involved dataset augmentation with 'negative' sequences that do not have RNA cleaving activity, which were RNA-ligating DNAzymes exhibiting opposite activity. These sequences are presented in DNAmoreDB, enabling the extraction of 188 sequences, which were designated as inactive. The progressive inclusion of these sequences into the dataset initially leads to a decrease in model accuracy, followed by a gradual enhancement of performance metrics during CV (Fig. 2). Notably, a 33% enhancement is observed at the final iteration, culminating in a final $Q^2$ value of 0.44. Therefore, it was demonstrated that inclusion of RNA-ligating DNAzymes as negative samples not only allows to account for data imbalance problem but also to increase model overall performance by reducing overestimation of DNAzyme $k_{obs}$.

#### Effect of experimental parameters

As previously indicated, the observed reaction $k_{obs}$ of DNAzyme depends strongly on specific experimental conditions (pH, cofactor type and concentration, temperature, NaCl etc.). To provide a more comprehensive characterization of the system, we have incorporated features pertaining to the composition of the buffer solution. These features have been transformed into a vector representation, with each element corresponding to a distinct component or parameter. A careful evaluation of parameters reveals their individual impacts on model performance. Combining buffer and sequence features significantly improves accuracy over merging sequence and cofactor descriptors (Table 3). This suggests a stronger correlation between buffer composition and the outcome, while cofactor attributes mainly determine DNAzyme activity as positive or negative. Integration of all three categories of descriptors results in a notable enhancement in the model accuracy during CV ($Q^2 = 0.62$). To address the issue of limited buffer diversity, we incorporated data regarding $k_{obs}$ dependence on various buffer parameters for sequences already present in the dataset, which led to further increase in performance to $Q^2 = 0.692$.

#### Top-performing model development

Based on the previous experiments, a dataset comprising 549 DNAzyme systems was compiled, consisting of total 349 unique (non-zero Levenshtein distance to each other) sequences and 71 parameters. Several models were tested on the ability to predict $k_{obs}$ on the final set of parameters, which are RF, XGBoost, LightGBM, and CatBoost, where100 samples (test set) allowed to tune model hyperparameters, whereas performance on 53 samples (validation set) show model performance on unseen data (see all models performance in Table S3).

All models exhibited consistent performance (Fig. 3a), displaying coefficient of determination values exceeding 0.89 in all cases. Notably, each model demonstrated

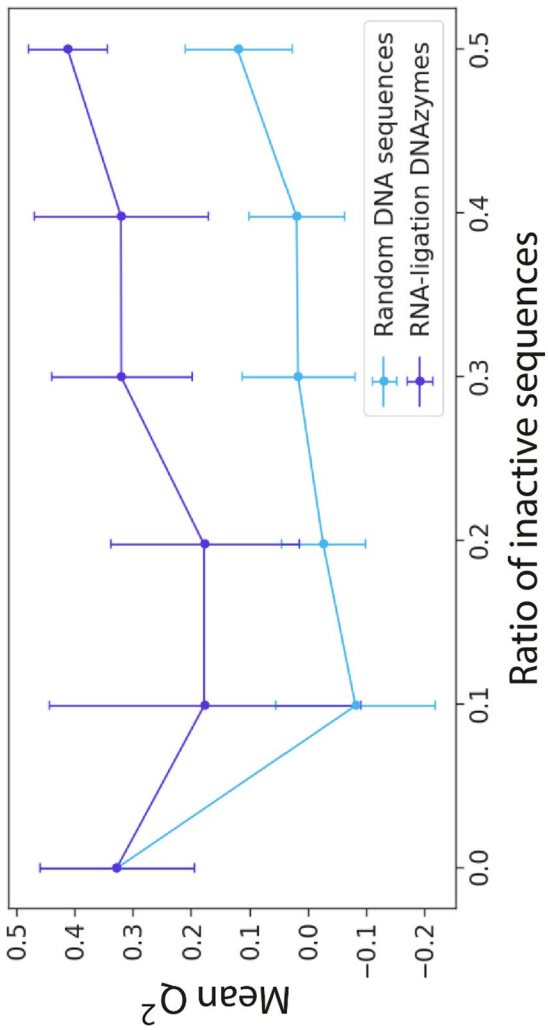


**Fig. 2** Effect of data balancing with addition of random and RNA-ligating sequences on averaged ML model performance

**Table 3** Effect of buffer- and cofactor-related descriptors on ML models performance

| Descriptors | $Q^2$ | $R^2_{test}$ | $RMSE_{test}$ |
|---|---|---|---|
| cofactor + sequence | 0.472 | 0.378 | 0.802 |
| conditions + sequence | 0.523 | 0.502 | 0.719 |
| conditions + sequence + cofactor | 0.617 | 0.472 | 0.740 |
| only sequence | 0.328 | 0.336 | 0.829 |

proficient capability in distinguishing between active and inactive samples, with the logarithm of $k_{obs}$ not exceeding −4 for inactive sequences, which shows model ability to detect inactive sequences. Among evaluated models, LightGBM model emerged as the most favorable, showcasing consistent accuracy across the spectrum of predicted values. In contrast, CatBoost and Random Forest models, although yielding relatively high metrics, displayed a propensity to overestimate the activity of DNAzymes with actual rate constants around 0.0001. At the same time, XGBoost tends to overfit leading to decrease of CV. Therefore, LightGBM model was selected for further optimization and final implementation due to its superior performance across all evaluation metrics ($Q^2 = 0.93$, $R^2_{test} = 0.97$, $R^2_{validation} = 0.96$) when utilizing all 71 parameters.

To remove parameters redundant and irrelevant for this task and increase model interpretability, recursive feature elimination (RFE) was implemented as feature selection method, where model performance plateaued at 24 parameters (Fig. 3b). Feature importance values analysis (Fig. 3c) has shown the superior importance of such experimental conditions as cofactor concentration, pH, ionic strength, and temperature, where the concentration of cofactor central to the process of hydrolysis, has the greatest effect on $k_{obs}$, which is consistent with experimental knowledge. Moreover, the model was able to determine the importance of such cofactor parameters as charge and electron affinity influencing nucleophilic attack on phosphodiester bond. AE-derived property-based descriptors related to the number of hydrogen bond donors and acceptors as well as correlated with GC content also were marked as important. At the same time, although improving overall model performance, k-mer based descriptors are not interpretable with feature importance values due to its intrinsic collective contribution to $k_{obs}$ value, whereas feature importance presumes each parameter contributes predicted value independently. Therefore, top-performing ML model predicting $lg(k_{obs})$ with RMSE = 0.52 e.g., capable of estimating $k_{obs}$ with an error of 1/3 order of magnitude in a wide range from 0.0001 to 1.7 $min^{-1}$ was established, optimized, and feature selected (Fig. 3c), where its interpretability and consistence with common knowledge was demonstrated.

#### SequenceCraft web resource

To ease the utilization of these findings by experimental scientists working on DNAzymes optimization and discovery, open source, and user-friendly web resource SequenceCraft (https://sequencecraft.aicidlab.itmo.ru/) was developed (see detailed description in ESI). This resource not only contains plenty of reference information about DNAzymes for those who are unfamiliar with the field, but also gives an access to the curated database of RNA-cleaving DNAzyme catalytic cores, as well as the predictive

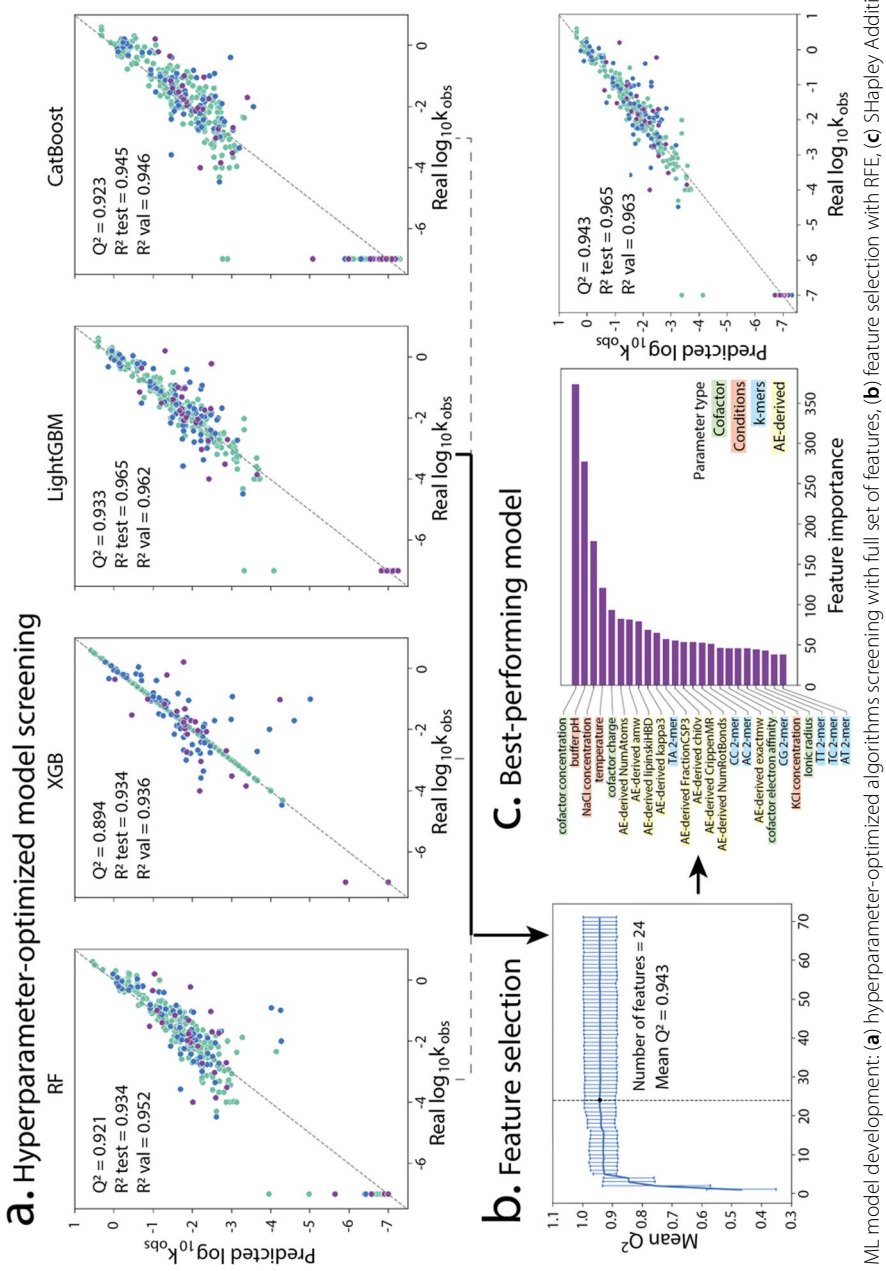


**Fig. 3** Best-performing ML model development: (**a**) hyperparameter-optimized algorithms screening with full set of features, (**b**) feature selection with RFE, (**c**) SHapley Additive exPlanations (SHAP) feature importance (left) and real vs. predicted plot (right) for best-performing LightGBM model

algorithm presuming sequence $k_{obs}$ based on sequence composition and experimental setup.

## Conclusion

Therefore, in this work, ML based open-source platform SequenceCraft was developed and deployed allowing experimental scientists to perform DNAzyme exploratory analysis via quantitative $k_{obs}$ estimation as well as statistical and clustering data analysis. This became possible with the development of unique curated database of 361 RNA-cleaving catalytic cores, which contains parameters characterizing catalytic core sequence composition, metal cofactor elemental properties, and experimental setup. Thorough sequence analysis allowed to ensure data completeness and construct the optimal set of parameters able to presume $k_{obs}$ values with high precision. Moreover, comprehensive comparison of existing sequence representation methods and its impact on $k_{obs}$ estimation revealed the superior performance of novel AE-derived property-based descriptors proposed in this work for the first time. To account for data imbalance problem common for every research at the intersection of chemistry and AI and limiting the utilization of novel predictive algorithms, RNA-ligating sequences were added as negative samples demonstrating increase in model performance in comparison with randomly generated sequences commonly used for these purposes. Further, a set of top-performing random forest and boosting-based ML was pre-optimized and screened, followed by feature selection and final hyperparameter optimization resulting in best-performing LightGBM model with performance $Q^2 = 0.93$.

Despite an effective algorithm for exploratory analysis of DNAzyme catalytic core activity was successfully developed in this paper, it is crucial also to highlight future challenges in this field to be addressed to achieve SNP level precise ML algorithms. Since some of the key parameters influencing the particular value of catalytic activity are absent in structured databases e.g., its dependence on dinucleotide cleavage site composition and SNPs not only for the most common catalytic cores, more mutation experimental data as well as the development of LLM-based reinforcement learning from human feedback (RLHF) algortihms should be implemented to account for insufficient data and absent parameters. At the same time, the lack of high-resolution X-ray diffraction structures as well as precise molecular mechanisms for the majority of DNAzymes limits the development of highly interpretable models able to give valuable insights on algorithm decision making process and propose novel mechanisms.

Overall, the findings regarding the possibility of DNAzyme exploratory analysis using ML presented in this paper make the first step towards SNP level precise catalytic activity predicting algortihms and promote the use of data-driven approaches by experimental scientists in the field of DNAzyme optimization and discovery.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-06019-7.

Additional file 1.

**Acknowledgements**
We thank Golovkin Ivan and Karemina Anastasiia for their help in data collection and conduction of statistical data analysis.

**Author contributions**
M.E. collected data and implemented ML pipeline. Y.D. and N.Sh. wrote the backend and frontend of the web service. N.S. developed a tool for autoencoder descriptors calculation and supervised the work. M.E. and N.S. wrote the main manuscript text. All authors reviewed the manuscript.

**Funding**
The work was financially supported by Russian Science Foundation no. 24–24-00546. The authors also thank Priority 2030 Federal Academic Leadership Program for infrastructure support.

**availability of data and materials**
The data extracted from DNAmoreDB are accessible in the original database (https://www.genesilico.pl/DNAmoreDB/dnazymes). ML models utilized in this study, along with the datasets employed for SequenceCraft is available on GitHub (https://github.com/GenerativeMolMachines/dnaZyme) as well as web service source code is available on GitHub (https://github.com/GenerativeMolMachines/dnaZymeWeb) to improve the reproducibility and transparency of the study.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 21 July 2024 Accepted: 17 December 2024
Published online: 06 January 2025

**References**

1. Taylor AI, Holliger P. On gene silencing by the X10–23 DNAzyme. Nat Chem. 2022;14:855–8.
2. Lake RJ, Yang Z, Zhang J, Lu Y. DNAzymes as activity-based sensors for metal ions: recent applications, demonstrated advantages, current challenges, and future directions. Acc Chem Res. 2019;52:3275–86.
3. Ali MM, Wolfe M, Tram K, Gu J, Filipe CDM, Li Y, Brennan JD. A DNAzyme-based colorimetric paper sensor for *Helicobacter pylori*. Angew Chem. 2019;131:10012–6.
4. Du M, Zheng J, Tian S, Liu Y, Zheng Z, Wang H, Xia J, Ji X, He Z. DNAzyme walker for homogeneous detection of enterovirus EV71 and CVB3. Anal Chem. 2021;93:5606–11.
5. Peng H, Newbigging AM, Wang Z, Tao J, Deng W, Le XC, Zhang H. DNAzyme-mediated assays for amplified detection of nucleic acids and proteins. Anal Chem. 2018;90:190–207.
6. Wang Y, Nguyen K, Spitale RC, Chaput JC. A biologically stable DNAzyme that efficiently silences gene expression in cells. Nat Chem. 2021;13:319–26.
7. Stoltenburg R, Reinemann C, Strehlitz B. SELEX—a (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomol Eng. 2007;24:381–403.
8. Kohlberger M, Gadermaier G. SELEX: critical factors and optimization strategies for successful aptamer selection. Biotechnol Appl Biochem. 2022;69:1771–92.
9. Chan L, Tram K, Gysbers R, Gu J, Li Y. Sequence mutation and structural alteration transform a noncatalytic DNA sequence into an efficient RNA-cleaving DNAzyme. J Mol Evol. 2015;81:245–53.
10. Ponce-Salvatierra A, Boccaletto P, Bujnicki JM. DNAmoreDB, a database of DNAzymes. Nucleic Acids Res. 2021;49:D76–81.
11. Mohammadi-Arani R, Javadi-Zarnaghi F, Boccaletto P, Bujnicki JM, Ponce-Salvatierra A. DNAzymeBuilder, a web application for *in situ* generation of RNA/DNA-cleaving deoxyribozymes. Nucleic Acids Res. 2022;50:W261–5.
12. Pine AC, Brooke GN, Marco A. A computational approach to identify efficient RNA cleaving 10–23 DNAzymes. NAR Genom Bioinf. 2023;5(1):lqac98.
13. Ma L, Kartik S, Liu B, Liu J. From general base to general acid catalysis in a sodium-specific DNAzyme by a guanine-to-adenine mutation. Nucleic Acids Res. 2019;47:8154–62.
14. Drevon D, Fursa SR, Malcolm AL. Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. Behav Modif. 2017;41:323–39.
15. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci. 2013;68:314–9.
16. Dong J, Yao Z-J, Zhang L, Luo F, Lin Q, Lu A-P, Chen AF, Cao D-S. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. J Cheminform. 2018;10:16.
17. Pereira F, Azevedo F, Carvalho Â, Ribeiro GF, Budde MW, Johansson B. Pydna: a simulation and documentation tool for DNA assembly strategies using python. BMC Bioinf. 2015;16:142.

18. Nguyen E. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *ArXiv*. 2023.
19. Peracchi A, Bonaccio M, Clerici M. A mutational analysis of the 8–17 deoxyribozyme core. J Mol Biol. 2005;352:783–94.
20. Akkaya UM, Kalkan H. Classification of DNA sequences with k-mers based vector representations. In: 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, pp. 1–5. 2021. https://doi.org/10.1109/ASYU52992.2021.9599084.
21. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15. https://doi.org/10.1093/nar/gkg595.
22. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*; 2022
23. Li Y, Breaker RR. Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2'-hydroxyl group. J Am Chem Soc. 1999;121:5364–72.

## Publisher's Note