Enhancing CRISPR-Cas9 On-Target Activity Prediction Employing CatBoost Machine Learning Models

Kien Trung Hoang
Faculty of Information Technology
VNU University of Engineering and Technology,
Vietnam National University, Hanoi
Hanoi, Vietnam
trungkienhoang1999@gmail.com

Chi Kim Hoang
Institute of Natural Products Chemistry,
Vietnam Academy of Science and Technology
Hanoi, Vietnam
chihoangkim@gmail.com

Diep Thi Hoang
Faculty of Information Technology
VNU University of Engineering and Technology,
Vietnam National University, Hanoi
Hanoi, Vietnam
diepht@vnu.edu.vn
*Corresponding author

2024 16th International Conference on Knowledge and System Engineering (KSE) | 979-8-3315-0940-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/KSE63888.2024.11063519

***Abstract*—The CRISPR-Cas9 system has revolutionized genome editing. Its effectiveness relies on selecting single guide RNAs (sgRNAs) that efficiently target the desired genomic location. Predicting the on-target activity of sgRNAs is crucial for optimizing CRISPR experiments and minimizing off-target effects. This paper develops CatBoost machine learning models to predict the on-target efficacy of CRISPR-Cas9 sgRNAs. We train the models on large benchmark datasets of sgRNAs targeting genes. An extensive set of sequence and genomic features are engineered as inputs. The CatBoost models outperform examined gradient-boosted regression trees and deep learning methods in predicting sgRNA activity. Feature importance analysis reveals the critical factors influencing sgRNA activity prediction. This work contributes to the field of precision genome engineering and has potential applications in biomedical research, agriculture, and biotechnology.**

***Index Terms*—CRISPR-Cas9, machine learning, sgRNA, on-target activity, genome editing**

I. Introduction

Genome editing, which involves the deletion, insertion, replacement, and alteration of DNA at specific sites within a living organism's genome, has emerged as a powerful tool for genetic engineering [5], [13], [14]. Among the various editing methods available, CRISPR-Cas9 has become the most widely adopted technique globally due to its versatility, accessibility, and effectiveness in modifying target genes across multiple species and biological contexts [2], [3].

The efficacy of CRISPR-Cas9 genome editing heavily depends on the selection of single guide RNAs (sgRNAs) that target the desired genomic locations. Predicting the on-target activity of sgRNAs is crucial for optimizing CRISPR experiments and minimizing off-target effects. Early methods for predicting on-target efficacy and off-target effects of CRISPR-Cas9 focused on calculating scores based on the positions and types of mismatches between the sgRNA and the target DNA sequence [10], [15]. However, these early approaches often overlooked the broader genomic context and were limited to basic sequence features [7].

To enhance prediction accuracy, advanced machine learning and deep learning models have been introduced for genome editing in both human and plant systems [18], [21]. These models incorporate a diverse set of features, such as specific characteristics of the genomic target, sgRNA thermodynamics, and pairwise similarity between the sgRNA and the genomic target. Machine learning has demonstrated considerable success in identifying gene cleavage sites, offering a promising strategy for reducing off-target effects in CRISPR-Cas9 gene editing [11], [21].

This study aims to develop a novel machine-learning model using the CatBoost algorithm [19] to predict the on-target performance of CRISPR-Cas9 sgRNAs. By utilizing large-scale datasets [11], [21] and integrating a comprehensive range of sequence and genomic features, we seek to enhance the accuracy of sgRNA efficacy predictions. The proposed model has the potential to improve CRISPR experiment design, accelerate research across fields such as biomedicine and biotechnology, and contribute to the advancement of precision genome engineering [16], [17].

II. Background

A. Problem statement

The problem at hand is an on-target prediction task for CRISPR-Cas9 sgRNA efficiency. The input data consists of the sgRNA sequence, target gene, and other relevant information. The output is a single score ranging from 0 to 1, representing the sgRNA's effectiveness, with scores closer to 1 indicating higher efficiency in causing knockout mutations of the target gene, and scores closer to 0 suggesting poor activity. Essentially, this problem can be framed as a regression task in machine learning, aiming to develop a model that accurately predicts the on-target efficiency of sgRNAs. The model's performance is evaluated using the Spearman correlation coefficient.

Describe dataset (Figure 1):

- Nucleotide composition: Includes position-specific features (nucleotide identity at each position) and position-independent features (total count of each nucleotide) in the sgRNA sequence.
- Dinucleotide composition: Considers adjacent nucleotide pairs, both position-specific and position-independent, similar to nucleotide composition.

- GC content: This feature represents the percentage of guanine and cytosine nucleotides in the sgRNA sequence.
- Thermodynamic properties: Based on predicted melting temperatures of the sgRNA sequence and its parts, including the entire sgRNA, 5' and 3' ends, and seed region.
- Target site location: Represented by amino acid cut position, measuring the distance from the protein-coding start to the sgRNA cut site within the gene.
- Protein conservation: Represented by percent peptide, likely indicating conservation of the targeted amino acid across related species.
- Flanking sequence: Considers nucleotide composition of two nucleotides upstream and downstream of the PAM, adjacent to the sgRNA target site.

The above data will be processed through machine learning algorithms with input: $(\mathbf{x}_i, y_i)_{i=1}^n$, I, α, L, s. And Output value is $F(\mathbf{x})$

	Sequence	Target gene	Percent Peptide	Amino Acid Cut position	Tm global_False	5mer_end_False	8mer_middle_False	5mer_start_False	GC count	score
0	AAAAAAAACACTGCAACAAG	CD5	72.87	360.0	57.383334	-36.413931	11.937531	-50.319302	6	0.083682
1	AAAAAGCAGCGTCAGTGGAT	CD5	84.21	416.0	64.634037	-40.944401	15.336802	-32.966948	9	0.184100
2	AAAACAACGGCCCAGGAGGG	CD5	52.23	258.0	70.519427	-26.959420	22.728950	-42.573078	12	0.188285
3	AAAAGGAAGATTGCTGATGA	CD45	41.95	474.0	59.262846	-38.864159	4.311995	-41.141740	7	0.581227
4	AAAAGTATCAGTGTGTATAG	THY1	76.54	124.0	53.340751	-54.115450	8.213311	-49.284841	6	0.677419

Fig. 1. Input data for prediction model.

B. Related work

Several prior studies have developed machine learning models to predict both the on-target efficacy and off-target effects of CRISPR-Cas9 sgRNAs. Doench et al. introduced the Rule Set 1 model, which predicted sgRNA on-target activity based on data from 1,841 sgRNAs targeting nine human and mouse genes [2]. This model utilized a support vector machine classifier with position-specific nucleotide features. The authors applied this model to design CRISPR libraries for humans and mice, named Avana and Asiago. In a subsequent study, Doench et al. expanded the training dataset to over 4,000 sgRNAs and developed an enhanced Rule Set 2 model using gradient-boosted regression trees (GBT) along with an expanded feature set [2]. This improved model outperformed Rule Set 1 and was employed to design the Brunello and Brie CRISPR libraries.

Doench et al. compared several machine learning algorithms for sgRNA activity prediction, including support vector machines (SVM) combined with logistic regression (used in Rule Set 1), SVM alone, L1 logistic regression, L1 linear regression, L2 linear regression, Random Forest, GBT, and SVM classification. Among these, GBT demonstrated the best performance and was adopted as the final model in the Azimuth tool. Wang et al. generated a much larger dataset of over 50,000 sgRNAs targeting 20,000 human genes for wild-type SpCas9, eSpCas9(1.1), and SpCas9-HF1 [21]. Using this dataset, they developed an optimized deep learning model called DeepHF, which integrates a recurrent neural network (RNN) with biological features [21]. In their study, they compared various machine learning approaches, including linear regression, L2-regularized linear regression (Ridge regression), XGBoost regression, multilayer perceptron (MLP), convolutional neural network (CNN), and RNN. The RNN model outperformed the other approaches and was selected as the final model for DeepHF.

Despite significant progress in modeling CRISPR on-target and off-target effects, opportunities remain to enhance predictive accuracy and evaluate models on additional sgRNA efficacy datasets. Building on these foundations, we investigate a new model that leverages large-scale training data to predict sgRNA activity.

III. Method

A. Typical Boosting Algorithms

Given an input matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ of n samples and m features, and a vector $\mathbf{y} \in \mathbb{R}^n$ of target values, the goal of gradient boosting is to approximate the underlying function $F(\mathbf{x})$ that maps input features $\mathbf{x}_i$ to target values y_i [23] as:

$$\hat{F}(\mathbf{X}) = \sum_{t=1}^{I} \alpha \hat{T}_t(\mathbf{X}) \tag{1}$$

where α is the learning rate, a regularization parameter controlling the influence of each tree $\hat{T}_t$ in the ensemble of I trees. For a loss function $L(\mathbf{y}, \hat{\mathbf{y}})$ measuring the difference between true values $\mathbf{y}$ and predictions $\hat{\mathbf{y}}$, each new tree $\hat{T}_t$ is learned to minimize [23]:

$$\hat{T}_t = \arg\min_{T_t} \mathcal{R} \left[-\frac{\partial L(\mathbf{y}, \hat{F}_{t-1}(\mathbf{X}))}{\partial \hat{F}_{t-1}(\mathbf{X})} - T_t(\mathbf{X}) \right] \tag{2}$$

where the derivative of the loss with respect to the ensemble output represents the prediction residuals of $\hat{F}(\mathbf{X})$ at the previous iteration, and T_t is the decision tree at the current iteration t. Each new tree compensates for the errors of the ensemble up to the previous iteration, analogous to gradient descent in function space [23].

XGBoost introduces a regularized objective:

$$L^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \Omega(T_t) \tag{3}$$

where L is a loss function and $\Omega(T_t) = \gamma J_t + \frac{1}{2}\lambda \|\mathbf{w}_t\|^2$ is a regularization term on the complexity of tree T_t, with J_t leaves and leaf weights $\mathbf{w}_t$. The parameters γ and λ control the regularization, leading to trees with fewer leaves and smoother weights that generalize better [23]. XGBoost also employs Newton boosting for faster convergence and a sparsity-aware split-finding algorithm to efficiently handle sparse data [12].

Gradient boosting employs a depth-first search method, resulting in asymmetric trees [1]. XGBoost uses a breadth-first search method, maintaining a constant tree depth across all branches, creating a balanced tree structure with equal depth in all branches [12]. On the other hand, CatBoost constructs oblivious trees, where the same split index is used across all branches at a given depth [19] (Figure 2). This approach helps reduce overfitting and improves the model's generalization.

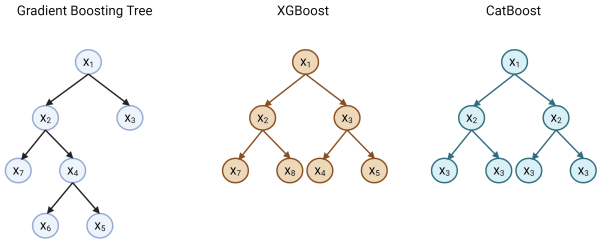

Fig. 2. Different tree structures and split indexes (shown inside each node) generated by Gradient Boosting, XGBoost and CatBoost.

B. Proposed model for on target prediction

This study utilized a dataset of experimentally validated CRISPR-Cas9 target sequences. The model developed is applicable to CRISPR-Cas9 across different species. Given the widely used sgRNA design rules, target sites from diverse organisms are likely impacted by similar genomic features considered in this research. Data pre-processing was performed to handle missing values. After selecting relevant attributes, feature extraction was conducted. The extracted features were used to train a CatBoost model for on-target prediction. The model was evaluated using Spearman correlation metrics. The following subsections detail the steps involved in the proposed model as depicted in Figure 3. Each step is described in detail, including the sequential process of data collection and model training.

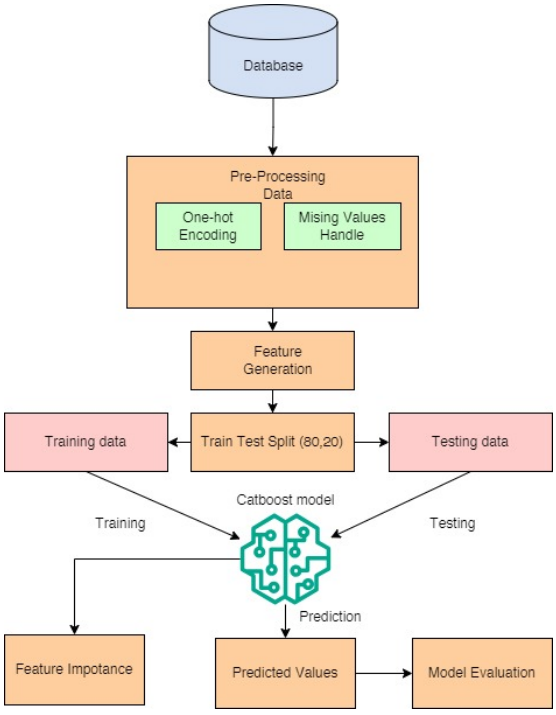

Fig. 3. Proposed model for predicting on-target activities of sgRNA.

Data Pre-Processing: The median filtering method was used to deal with missing data values during the pre-processing phase of data analysis. By considering the local neighborhood of values, median filtering provides a robust estimate for the missing data points while minimizing the influence of outliers. One-hot encoding is used to convert categorical variables with multiple categories into binary vectors. In the case of sgRNA sequences, each position in the 20-nt sequence can be one of four nucleotides (A, C, G, or T).

TABLE I
LIST OF ALL THE FEATURES

	Features derived from nucleotide contents
1	Position nucleotide
2	GC content
3	Number nucleotide
4	Melting temperature
5	Secondary structure

Feature Generation: Once the sgRNA data was cleaned and transformed, the next crucial step was to generate informative features from the sgRNA sequences (Table I). These features aim to capture the inherent characteristics and properties of the sgRNAs that may influence their efficacy and specificity in the genome editing process.

CatBoost Algorithm Architecture: The CatBoost regressor, an advanced gradient boosting algorithm, is used for supervised learning (Algorithm 1). CatBoost effectively handles categorical features like PAM sequences and Target genes without requiring extensive encoding.

Algorithm 1 Pseudocode of CatBoost for on-target prediction

Require: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, I, α, L, s
$\sigma_r \leftarrow$ random permutation of $[1, n]$ for $r = 0, \ldots, s$
$M_0(i) \leftarrow 0$ for $i = 1, \ldots, n$
for $j \leftarrow 1$ to $\lceil \log_2 n \rceil$ **do**
  $M_{r,j}(i) \leftarrow 0$ for $r = 1, \ldots, s$, $i = 1, \ldots, 2^{j+1}$
end for
for $t \leftarrow 1$ to I **do**
  $T_t, \{M_r\}_{r=1}^{s} \leftarrow \text{BuildTree}(\{M_r\}_{r=1}^{s}, \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \alpha, L, \{\sigma_r\}_{r=1}^{s})$
  $leaf_0(i) \leftarrow \text{GetLeaf}(\mathbf{x}_i, T_t, \sigma_0)$ for $i = 1, \ldots, n$
  $\text{grad}_0 \leftarrow \text{CalcGradient}(L, M_0, \mathbf{y})$
  for each leaf j in T_t **do**
    $b_j^t \leftarrow -\text{avg}(\text{grad}_0(i) \mid i : leaf_0(i) = j)$
  end for
  $M_0(i) \leftarrow M_0(i) + \alpha b^t_{leaf_0(i)}$ for $i = 1, \ldots, n$
end for
return $F(\mathbf{x}) = \sum_{t=1}^{I} \sum_j \alpha b_j^t \mathbf{1}_{\{\text{GetLeaf}(\mathbf{x}, T_t) = j\}}$

- $(\mathbf{x}_i, y_i)_{i=1}^{n}$: Training dataset with n samples, where $\mathbf{x}_i$ is the feature vector and y_i is the on-target score of the i-th sample.
- I: Number of decision trees in the boosting process.
- α: Learning rate, a constant that adjusts learning speed and each tree's influence on the model.
- L: Loss function used to evaluate and optimize the model.

- s: Number of random permutations of data samples for auxiliary models.
- σ_r: r-th random permutation of sample indices, where $r = 0, 1, \ldots, s$.
- M_0: Main model, initialized at the beginning and updated each boosting step.
- $\{M_r\}_{r=1}^{s}$: r-th auxiliary model at step j during tree construction.
- T_t: Decision tree built at the t-th boosting step.
- $leaf_0(i)$: Index of the leaf containing the i-th sample in tree T_t using permutation σ_0.
- grad_0: Gradient of the loss function with respect to current predictions of M_0.
- b_j^t: Value assigned to the j-th leaf of T_t, calculated as the negative average of gradients of samples in this leaf.
- BuildTree: Function that builds a new decision tree and updates auxiliary models. Outputs T_t (the new tree) and updated models $\{M_r\}$.
- GetLeaf$(\mathbf{x}_i, T_t, \sigma_r)$: Function returning the leaf index for sample $\mathbf{x}_i$ in tree T_t with permutation σ_r.
- CalcGradient$(L, M_0, \mathbf{y})$: Function that calculates the gradient of loss L relative to predictions M_0 and labels $\mathbf{y}$.
- $F(\mathbf{x})$: Final prediction model, which is a weighted sum of trees. For each new sample $\mathbf{x}$, $F(\mathbf{x})$ finds the corresponding leaf in each tree and sums the weighted values.

CatBoost efficiently handles both numerical and categorical data. Unlike most machine learning algorithms that require manual label encoding for categorical features, CatBoost automatically performs optimized label encoding. It analyzes the relationship between categorical features and the target variable, simplifying data preparation and providing a more efficient approach to handling categorical data compared to traditional methods.

IV. Experiment

A. Dataset

We utilized two datasets in this study to train and evaluate the CatBoost model for predicting sgRNA activity.

- Dataset from Doench et al. (2016) [11]: This dataset consists of 4,997 sgRNAs targeting 17 genes in human cells, along with their measured indel frequencies. The sgRNAs span a wide range of on-target activities. We selected this dataset because it is widely considered a benchmark in the field, being one of the most cited and utilized datasets for CRISPR sgRNA efficacy prediction.
- Dataset from Wang et al. (2019) [21]: This dataset contains over 50,000 sgRNAs targeting 20,000 human genes for three Cas9 variants - wild-type SpCas9 (WT-SpCas9), enhanced SpCas9 (eSpCas9(1.1)), and SpCas9-HF1. The indel frequencies were determined by genome-scale screening in human cells. We chose this dataset because it is currently the largest available dataset for CRISPR sgRNA activity prediction [24].

These two datasets together provide a comprehensive collection of sgRNA sequences and their on-target activities across different organisms and Cas9 variants (Table II). The Doench 2016 and Wang 2019 datasets allow the model to learn sequence features associated with sgRNA activity. The processed datasets were then randomly split into training (64 %), validation (16 %), and test (20 %) sets using stratified sampling to ensure balanced representation of different activity levels.

TABLE II
SUMMARY OF THE DATASETS

Dataset	*Sub-dataset*	*No. of Target genes*	*No. of sgRNA*
Doench et al.		17	4,997
Wang et al.	WT-SpCas9	18	55,604
	eSpCas9(1.1)	18	58,617
	SpCas9-HF1	18	56,888

B. Model training and evaluation

CatBoost was employed as the primary machine learning algorithm for predicting sgRNA activity. The training process involved optimizing the hyperparameters of the CatBoost model using Bayesian optimization [14]. The objective function for the optimization was the Spearman correlation coefficient between the predicted and observed activity levels on the validation set [11]. To establish baselines for comparison, we trained and evaluated several other machine learning models on the same dataset, including the gradient-boosted regression trees used in the Azimuth project [11] and RNN employed in the DeepHF project [21]. For the gradient-boosted regression trees, we implemented the model using the same features and training procedure as described in the Azimuth project [11]. The hyperparameters were optimized using a grid search approach. For the RNN model, we followed the architecture and training methodology outlined in the DeepHF [21] project. The sgRNA sequences were encoded using one-hot encoding, and the RNN was trained using the Adam optimizer with dropout regularization. The model performance was assessed using the same evaluation metrics as the CatBoost model. Cross-validation was employed to assess the robustness and generalization ability of the models [10]. The dataset was divided into k-folds, and the models were trained and evaluated k times (k=10), each time using a different fold as the validation set. The average performance across the folds was reported to provide a more reliable estimate of the models' performance.

C. Feature importance analysis

To gain insights into the relative importance of the input features, feature importance analysis was conducted employing the trained CatBoost model [22]. The feature importance values were to identify the most informative features for predicting sgRNA activity (Table III). The scores in Table III represent the PredictionValuesChange importance, which measures the cumulative change in the model's predictions when a feature is used for splitting [19]. This metric quantifies how much each feature contributes to the model's predictive power across all trees in the ensemble [20].

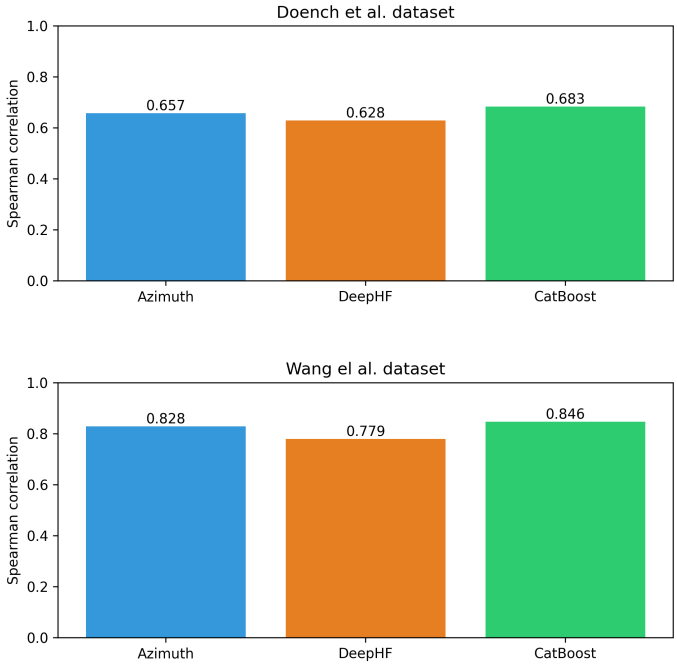

Fig. 4. Spearman correlation of three methods performance.

TABLE III
FEATURE IMPORTANCE

Feature	***Score***
Percent Peptide	10.94
5mer_end_False	6.39
8mer_middle_False	6.04
Amino Acid Cut position	5.90
5mer_start_False	4.81
Tm global_False	4.20
Target gene	2.76
G_23	2.71
A	2.47

D. Results and discussion

The CatBoost model demonstrated superior performance in predicting sgRNA on-target activity compared to examined methods, as evidenced by the higher Spearman correlation coefficients shown in Figure 4. This improvement in predictive accuracy can be attributed to CatBoost's ability to handle categorical features effectively and handle complex, non-linear relationships in the data robustly. CatBoost's superior performance in handling categorical features has been demonstrated in several previous studies. For instance, Hancock and Khoshgoftaar showed that CatBoost outperformed other gradient boosting algorithms when dealing with categorical data in various domains [22]. This efficient handling of categorical features, combined with other advanced techniques, makes CatBoost effective for CRISPR-Cas9 sgRNA activity prediction. The model's decision tree structure enables it to capture non-linear interactions, such as the interdependence of specific nucleotides and their positions within the sgRNA. Through iterative training, CatBoost can detect intricate patterns that might elude human recognition, like the impact of particular nucleotide combinations at different positions. The algorithm's contextual learning ability allows it to understand how various characteristics interact under different conditions, for instance, how GC content's influence may vary based on the cleavage position or cell type. CatBoost's robust overfitting prevention mechanisms enable it to discern true underlying relationships even in noisy experimental data, a common challenge in CRISPR-Cas9 datasets. Feature importance analysis revealed key insights into the factors influencing sgRNA efficacy. Interestingly, the top features identified by our CatBoost model show a high degree of similarity with those reported in the Azimuth study [11], with approximately 80% overlap. This consistency across different methodologies reinforces the significance of these features that determine sgRNA activity. The most important features identified include: Percent Peptide, Sequence composition, Amino Acid Cut position, Thermodynamic properties. Moreover, the identification of additional features by our CatBoost model underscores the potential of this approach to uncover new insights into the factors governing sgRNA efficacy. These findings can guide the development of more effective sgRNA design strategies and contribute to our understanding of the underlying mechanisms of CRISPR-Cas9 genome editing.

CONCLUSIONS

In this study, we developed machine learning models employing the CatBoost algorithm for predicting the on-target activity of CRISPR-Cas9 sgRNAs. By leveraging large-scale datasets encompassing sgRNAs targeting diverse genes across multiple species, our model learned to capture the complex relationships between sgRNA sequence, genomic context, and editing efficiency. The CatBoost algorithm's ability to handle categorical features and its robust performance make it well-suited for this prediction task. Our results demonstrate that the CatBoost model outperforms existing state-of-the-art methods, including gradient-boosted regression trees and recurrent neural networks, in terms of Spearman correlation. Feature importance analysis confirmed that position-specific nucleotide composition, thermodynamic stability, and genomic context are among the most informative features for predicting sgRNA efficacy. By enabling the selection of highly effective sgRNAs, our CatBoost model can facilitate more efficient and specific genome editing.

REFERENCES

[1] FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001, 1189-1232.
[2] FRIEDMAN, Jerome H.; MEULMAN, Jacqueline J. Multiple additive regression trees with application in epidemiology. Statistics in medicine, 2003, 22.9: 1365-1381.
[3] TERNS, Michael P.; TERNS, Rebecca M. CRISPR-based adaptive immune systems. Current opinion in microbiology, 2011, 14.3: 321-327.
[4] SNOEK, Jasper; LAROCHELLE, Hugo; ADAMS, Ryan P. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 2012, 25.

[5] SANDER, Jeffry D.; JOUNG, J. Keith. CRISPR-Cas systems for editing, regulating and targeting genomes. Nature biotechnology, 2014, 32.4: 347-355.
[6] ZHANG, Hui, et al. The CRISPR/C as9 system produces specific and homozygous targeted gene editing in rice in one generation. Plant biotechnology journal, 2014, 12.6: 797-807.
[7] SANJANA, Neville E.; SHALEM, Ophir; ZHANG, Feng. Improved vectors and genome-wide libraries for CRISPR screening. Nature methods, 2014, 11.8: 783-784.
[8] WESTRA, Edze R.; BUCKLING, Angus; FINERAN, Peter C. CRISPR–Cas systems: beyond adaptive immunity. Nature Reviews Microbiology, 2014, 12.5: 317-326.
[9] BORTESI, Luisa; FISCHER, Rainer. The CRISPR/Cas9 system for plant genome editing and beyond. Biotechnology advances, 2015, 33.1: 41-52.
[10] HAEUSSLER, Maximilian, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome biology, 2016, 17: 1-12.
[11] DOENCH, John G., et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology, 2016, 34.2: 184-191.
[12] CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.
[13] MA, Hong, et al. Correction of a pathogenic gene mutation in human embryos. Nature, 2017, 548.7668: 413-419.
[14] MUSUNURU, Kiran. Genome editing: the recent history and perspective in cardiovascular diseases. Journal of the American College of Cardiology, 2017, 70.22: 2808-2821.
[15] XU, Xiaojun; DUAN, Dongsheng; CHEN, Shi-Jie. CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment. Scientific reports, 2017, 7.1: 143.
[16] SOVOVÁ, Tereza, et al. Genome editing with engineered nucleases in economically important animals and plants: state of the art in the research pipeline. Current issues in molecular biology, 2017, 21.1: 41-62.
[17] LISTGARTEN, Jennifer, et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nature biomedical engineering, 2018, 2.1: 38-47.
[18] CHUAI, Guohui, et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome biology, 2018, 19: 1-18.
[19] DOROGUSH, Anna Veronika; ERSHOV, Vasily; GULIN, Andrey. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363, 2018.
[20] Prokhorenkova, Liudmila, et al. "CatBoost: unbiased boosting with categorical features." Advances in neural information processing systems 31 (2018).
[21] WANG, Daqi, et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nature communications, 2019, 10.1: 4284..
[22] HANCOCK, John T.; KHOSHGOFTAAR, Taghi M. CatBoost for big data: an interdisciplinary review. Journal of big data, 2020, 7.1: 94.
[23] BOLDINI, Davide, et al. Practical guidelines for the use of gradient boosting for molecular property prediction. Journal of Cheminformatics, 2023, 15.1: 73.
[24] Sherkatghanad, Zeinab, et al. "Using traditional machine learning and deep learning methods for on-and off-target prediction in CRISPR/Cas9: a review." Briefings in Bioinformatics 24.3 (2023): bbad131.