Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

Navigating the CRISPR-Cas9 Frontier: AI-Enabled off-target prediction and sgRNA Design for Unprecedented Precision

Haoyu Sun *

Department of Computer Science, University of Hong Kong, Hong Kong, China

* Corresponding Author Email: u3585604@connect.hku.hk

Abstract. CRISPR-Cas9, composed of Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein 9, is a pivotal tool for precise genetic manipulation with diverse biomedical applications. In this system, the scientists use single guide RNA (sgRNA) synthesized from tracrRNA and crRNA to lead the Cas9 protein to target specific gene locations, thereby achieving gene editing. Nonetheless, drawbacks like single guide RNA's limited efficiency led to frequent base mismatch and off-target effects, which hampers CRISPR-Cas9's potential. Under these circumstances, entering machine learning, adept at adapting to variations and handling intricate datasets, is a viable avenue for optimizing CRISPR-Cas9's guide RNA by rectifying these limitations. Nevertheless, machine learning is not exempt from limitations. Within this framework, this paper presents a succinct overview of the challenges linked to sgRNA's efficiency issues. It then outlines existing mechanisms in machine learning and assesses the efficacy of machine learning in enhancing sgRNA design to improve CRISPR/Cas9 sgRNA specificity. Additionally, this paper scrutinizes notable restrictions and suggestions of machine learning in the quest for superior sgRNAs to guide future research.

Keywords: CRISPR-Cas9; Machine learning; sgRNA design; efficient sgRNA.

1. Introduction

In the year of 2012, a significant breakthrough occurred when scientists affiliated with the University of California, Berkeley, unveiled CRISPR-Cas9's capabilities.[1] Their efficient gene editing approach showcased considerable promise for applications in plant genetic manipulation and the treatment of animal diseases. Today, CRISPR has set off an international trend of gene editing, and more advanced technologies have been developed. [2].

The current CRISPR-Cas9 gene editing process compasses 5 key steps: (1), the design of specific sgRNA, synthesized from tracrRNA and crRNA, direct Cas9 precisely to the target gene's precise position; (2), the amalgamation of gRNA with Cas9, leading to the creation of the sgRNA-Cas9 complex; (3), the complex scans the DNA for foreign sequences and identifies PAM motifs, targeting nearby gene sequences; (4), after attaching to the target DNA, the complex formed by sgRNA and Cas9 triggers the cleavage of the sequence through the action of the RuvC and HNH nuclease domains, leading to the formation of a DNA double-strand destruction; (5), the desired DNA repair template is introduced and connected to the cleaved target DNA's ends through homologous directed repair (HDR).[3] This process facilitates precise DNA edits like insertions or replacements during repair.

Compared with the other generations of gene editing techniques, a significant benefit of CRISPR-Cas9 lies in its notably reduced effort required for the design and synthesis of sgRNA when contrasted with the construction process of DNA recognition modules in other two technologies.[4] Furthermore, the specificity and cutting efficiency of different sgRNAs will directly influence the accuracy and efficiency of gene editing.[5] Therefore, designing optimal sgRNA is a crucial contribution to improving CRISPR-Cas9 systems.

However, how to design efficient sgRNA is still a focus that puzzles scientists. Hence, overcoming sgRNA design complexities demands advanced computational tools. In recent times, machine learning has offered novel approaches for addressing CRISPR/Cas9 challenges by utilizing past

knowledge and statistics, algorithms to create and update models, yielding precise analytical results. It is extensively employed in CRISPR/Cas9 systems to enhance efficient sgRNA sequence design [6] and predict sgRNA activity [2].

The obstacles to enhancing the performance of sgRNA through machine learning revolve around three primary domains:

Off-target prediction and elimination: Designing sgRNA requires preventing off-target effects, which manifest when the sgRNA-Cas9 complex fails to precisely locate the target gene sequence, leading to unintended alterations in non-target DNA regions [5]. These non-target regions are likely to be genomic sequences similar to the target. Unintended genetic mutations arising from off-target effects can disrupt cellular function, an organism's physiological processes, and potentially contribute to diseases. Mitigating these off-target effects holds utmost importance when employing CRISPR-Cas9 technology, particularly in clinical therapy.

Enhancing On-Target Activity: On-target activity refers to sgRNA's precision in initiating gene edits at the target site.[7] The high on-target activity involves considering complex factors like guiding sequence integrity, correspondence level, and genomic context.

Balancing Specificity and Editing Efficiency: The creation of effective sgRNA requires balancing specificity and editing efficiency [8]. Highly specific sgRNAs minimize off-target effects but have lower editing efficiency due to imperfect alignment with the guiding sequence and target site. Conversely, less specific sgRNAs may edit various target sites, raising the off-target risk.

Drawing from an elucidation of how machine learning is employed within the CRISPR-Cas9 framework to enhance the design of efficient sgRNAs, this analysis assesses the effectiveness of machine learning across critical optimization methodologies for sgRNA design. The significant defects are summarized to reference the progressive research in related fields.

2. Mechanisms of Machine learning in optimizing sgRNA design

Incorporating machine learning into the optimization of CRISPR-Cas9 systems, this section delineates the process of utilizing machine learning to enhance the design of efficient sgRNA in four distinct components. Research on how to use machine learning to advance the performance of sgRNA can start from several different angles. It is the primary task of research, but also, to a large extent, led to the differentiation and refinement of the current sgRNA-related research. As for the research tendency, the available experimental results list many potential research priorities:

1. Basic sequence features were studied to anticipate the off-target of sgRNAs, and a scoring system to rank the performance of sgRNA [9].

2. Methods such as evaluating the likelihood of genomic sites being cleaved by a particular sgRNA play a crucial role in understanding the importance of different factors when designing sgRNAs. [10]

3. Study the interaction between multiple Sgrnas and design pairs of Sgrnas to improve editing efficiency. [11].

2.1. Prior data

Prior data refers to the information or data that has been obtained prior to the analysis, modeling, or prediction. When designing sgRNA, these data are usually selected from NCBI RefSeq, *Genebank* database, and FC-RES dataset [3] according to research tendencies. These are useful for experiments, such as sgRNA cutting frequency and cutting site set [9]. Prior data can help researchers limit the model's search space, thereby improving the efficiency and accuracy of the model's analysis of sgRNA-specific features.

2.2. Model Construction

Depending on the selected features, different machine learning patterns and model construction will have a complex impact on the computational outcome. Therefore, choosing a model suitable for studying a given feature is crucial. Model selection determines the calculation's efficiency and the results' interpretability. Existing relevant algorithms suitable for sgRNA design include but are not limited to logistic regression [8], random forest [12], support vector machines and neural networks [13]. In addition, improvement and data comparison based on previous experience are essential. Through model comparison and integration [14], model optimization is also an essential direction in the field of sgRNA design to improve the interpretability and increase the accuracy of predicting sgRNA performance.

2.3. Conclusion Cross-validation and evaluation of clinical trials

After the calculation results are available, the cross-validation method is required to conduct a preliminary evaluation of algorithm performance to determine the model's generalization ability [8]. As an example, the sgRNA generated within the model demonstrates both on-target effectiveness and an unanticipated influence on newly encountered target data. Subsequently, it becomes imperative to initiate clinical trials in organisms like mice for further validation [15]. Although CAD can theoretically provide a set of sgRNA sequences with good predictive performance, in practice, the effect in biological systems can be affected by various factors.

This section will use specific examples to demonstrate the effectiveness of different machine learning principles and tools in optimizing sgRNA designs in different contexts for different research fields.

3. Case studies of ML boosted CRISPR editing efficiency

3.1. Application of Machine Learning in Predicting Off-Target Effects, Case - 1: Supervised Learning

Scientists from the Indian Institute of Technology utilized machine learning to create a computational model capable of precisely forecasting potential off-target locations in the body susceptible to interference. This algorithmic tool, focuses on the utilization of sgRNA (single-guide RNA), and leverages fundamental sequence properties, including nucleotide accessibility, mismatch count, GC content, and site-specific nucleotide conservation. Unlike those involving specific species or diseases, this study centers on pure algorithm development without wet experiments [16].

Utilizing various classifiers like gradient lift regression trees (xgboost), perceptrons, logistic regression, random forests and support vector machines, the study employed supervised learning techniques. 4-fold cross-validation for data partition has been applied here, with iterative adjustments of classifier hyperparameters. The best-performing classifier was the gradient lifting regression tree (xgboost), as determined by testing results. Feature importance analysis incorporated SHAP values and other techniques, ultimately identifying 18 significant features. This classifier offers precise predictions of off-target effects, providing valuable guidance for gene editing strategies.

The dataset used here encompassed 19 targets, with 6337 experimentally identified positive mismatch sites and 7040 experimentally unidentified negative mismatch sites. Positive mismatch sites and target data were obtained from GUIDE-seq and SITE-Seq studies, while negative mismatches were derived from the CRISPcutweb server. The test dataset consisted of 2,877 positive mismatch sites, 4,010 negative mismatches, and nine targets.

Employing supervised learning, models were constructed using the mentioned classifiers. The data segmentation process encompassed 4-fold cross-validation and the fine-tuning of hyperparameters. During training, the model was practiced on three folds, validating accuracy on the remaining fold. After multiple iterations, the xgboost model with optimized hyperparameters emerged as the final model, achieving up to 91.49% accuracy in testing.

The identification of crucial sequence descriptors employed by the selected model's "tree" mechanism followed.

1. Employing supervised learning, models were constructed using the mentioned classifiers. The data segmentation process encompassed 4-fold cross-validation and the fine-tuning of hyperparameters. During training, the model was practiced on three folds, validating accuracy on the remaining fold. After multiple iterations, the xgboost model with optimized hyperparameters emerged as the final model, achieving up to 91.49% accuracy in testing.

2. The identification of crucial sequence descriptors employed by the selected model's "tree" mechanism followed.

3. The analysis encompassed the assessment of feature importance within the xgboost classifier, employing metrics such as SHAP values, information gain, coverage rate, and frequency. Gain indicates feature-enhanced prediction accuracy in tree branches, coverage reflects observations, and frequency indicates feature usage. SHAP values, consistent and accurate, were considered alongside other metrics. Based on the SHAP ranking, 18 highly significant features were identified for influencing model predictions, forming a scoring mechanism.

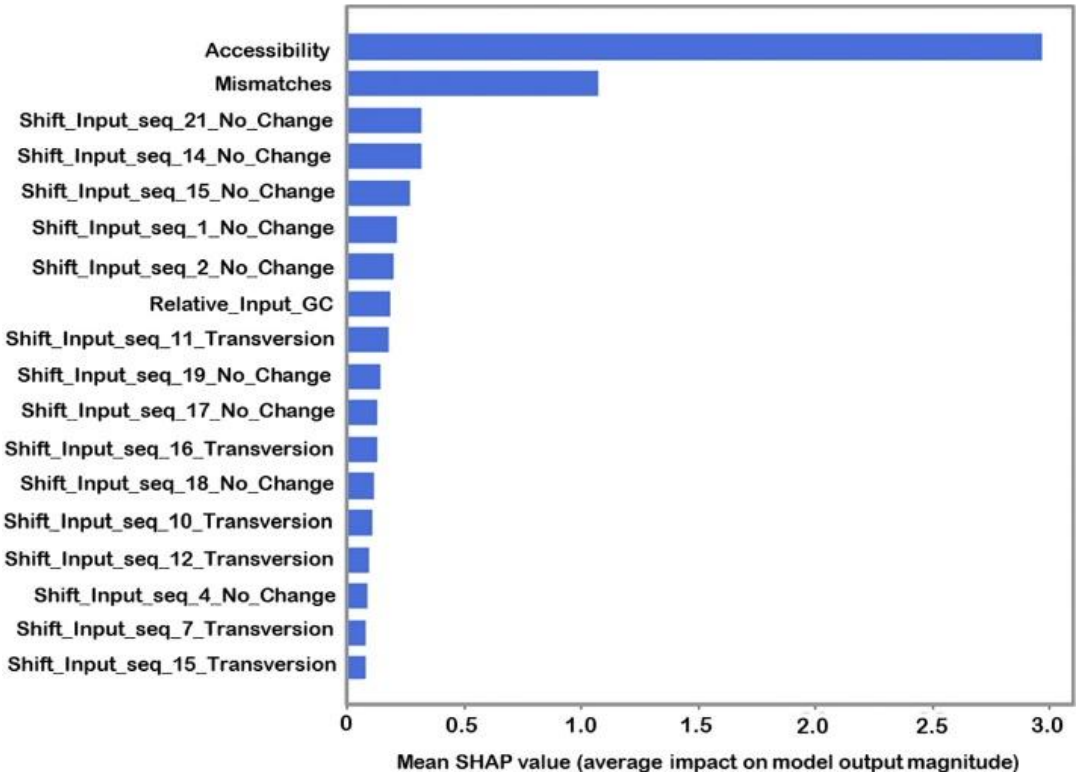

Figure 1. Sequence characteristics that make a substantial contribution to the overall effectiveness of the predictive model. Implementation of the global mean method (|Tree SHAP|) within the prediction model.

The model highlights the significance of GC content in predicting positive off-target sites and its influence on sgRNA off-target effects. Among 18 features assessed, target DNA accessibility is the primary predictor, succeeded by off-target site mismatch count (Figure 1). The model aligns with the CRISPR/Cas9 "seed" model, where "seed" region mismatches hinder sgRNA-DNA binding, impairing gene-editing. Consistency exists between positive off-target nucleotides and sgRNA target sequences within the "seed" region, except for one position. This collective insight unveils the intricate factors governing sgRNA behavior, demonstrating machine learning's potential to decipher these complexities for more precise genome editing.

This model predicts positive off-target site characteristics with 91.49% accuracy, resonating with experimental outcomes. Henceforth, this machine learning model can be seamlessly incorporated into

sgRNA design utilities, augmenting the precision of sgRNA predictions for particular genomic sites within specific cell lines, with a focus on mitigating off-target effects.

3.2. Application of Machine Learning in Predicting Off-Target Effects, Case - 2: Utilizing Random Forest-Based ML to Assist in CRISPRi Design of Multi-Coccal sgRNA

A study published in ACS Synthetic Biology, authored by Tessa Dallo et al., introduced the application of CRISPRi (Clustered Regularly Interspaced Short Palindromic Repeats) for the first time in polycoccus. The innovative interference method laid the foundation for gene editing of this microorganism. Moreover, the researchers further used random forest machine learning methods to predict the effectiveness of gRNA for 76 7002 strains cultured to wet experiments, thereby showcasing the potential of machine learning in enhancing gRNA design for the regulation of specific gene expression [17].

In this study, scientists employ machine learning together with high-density gRNA tiling and correlation analysis, to delve into the gRNA design principles within CRISPRi for Synchronous sp. PCC 7002. Generating a multitude of gRNAs through tiling, which enables a meticulous examination of design attributes for enhancing CRISPRi efficiency. Notably, the focus on strain 7002 stems from its rapid growth and robustness in the face of varying conditions like light, salt, and temperature, positioning it as a prospective genetic platform for biofuel production. Assessing gRNA's impact involves cultivating cultures with and without dCas9 and gRNA inducers. Additionally, correlation analysis pinpoints gRNA design guidelines, while machine learning forecasts gRNA performance. The study's machine learning component encompasses five specific regression models: XG Boost, Random Forest, decision tree, k-nearest neighbours, and linear regression with ridge L2 regularization.

Three parental strains were used here: Polycoccus PCC 7002, 7002-RBCP-YPET, and 7002-P2579-YPET. dCas9, and the combination of three strains into the genome of the parent strain by transforming the linearized pCas2C plasmid. The design of gRNA was accomplished using Benchling software, with a focus on target and mismatch scores to select suitable candidate sequences. aTc (isoleucine) -induced gRNA expression plasmid can detect the transformed strain and gene cloning constructed plasmid. Then, the guide RNA sequences were phosphorylated, annealed, and transformed. The verified transformants are then analyzed and validated.

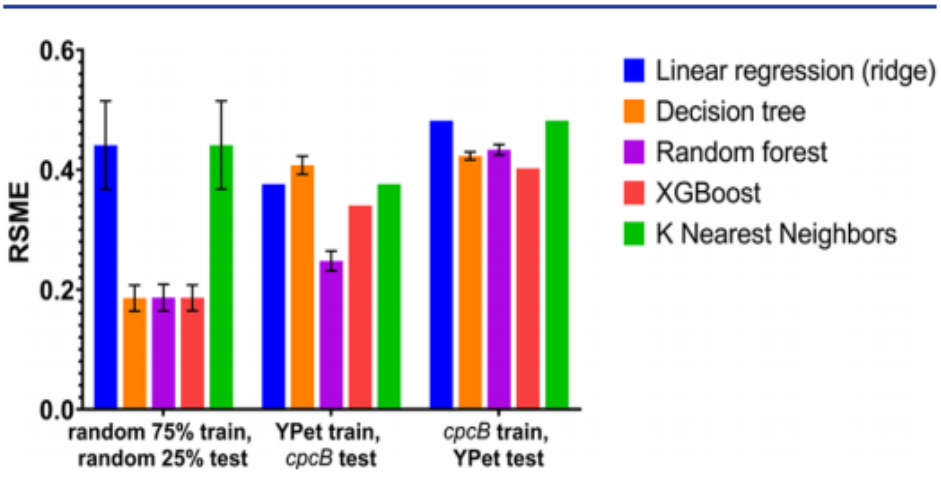

Figure 2. Machine Learning Model Performance in Predicting gRNA Efficiency. Three distinguished data sets have different root mean square errors (RMSE). The Error bars on the graph depict the standard deviation calculated from 10 separate trials.

Algorithm and Feature Selection This section includes the selection of five machine learning algorithms using the sci-kit-learn tool. Furthermore, there are various features to capture the relationship between features and results using data, including feature correlation.

Model Optimization and Evaluation The data processed with aTc includes training and test sets, and each model undergoes hyperparameter optimization through Bayes optimization. Model performance is evaluated by applying it to the test set, and the assessment entails calculating the root mean square error (RMSE) (Figure 2). Multiple experiments are conducted using randomized data partitioning.

Model Validation and Results Researchers employ ten-fold cross-validation to validate the trained model across diverse datasets, assess the consistency of its performance, and ascertain the deterministic or random nature of the obtained results.

In the realm of CRISPRi, machine learning helps optimize inhibition effectiveness by selecting template chain gRNA sequences close to the start codon, encompassing CGG or GGG PAM sites. This strategic selection mitigates off-target sequences within the genome. Integrating high-density gRNA datasets and machine learning techniques represents a pivotal advancement, enhancing the capacity to forecast gRNA efficiency in CRISPRi applications accurately. Notably, the efficacy of gRNA prediction for strain 7002 is particularly pronounced with the random forest model, exhibiting superior performance characterized by a minimal error rate during both the training and testing phases, which emphasizes the efficacy of machine learning in effectively gauging gRNA effects in a specific context. Furthermore, the convergence of machine learning models provides valuable insights into the evolutionary conservation of CRISPR targeting mechanisms across diverse species, thereby furnishing guidance for the regulation of gene expression.

This research stands as a prime example of using machine learning for prediction and improvement in the realm of CRISPR-Cas9-related gene editing methods, particularly when applied to strains grown through wet lab experiments. It underscores the significant role played by machine learning in aiding the design of sgRNAs for CRISPRi applications.

3.3. Application of Machine Learning in Predicting Off-Target Effects, Case - 3: DeepCRISPR, a Deep learning framework for predicting sgRNA efficiency

A pioneering effort led by Guo Hui Chuai and their team at Tongji University has given rise to an innovative computational platform known as *DeepCRISPR.* This platform seamlessly integrates on-target and off-target predictions within a deep learning framework while also employing a digitally driven approach to automatically discern sequences and epigenetic features with the potential to impact sgRNA knockout efficiency. This contribution yields a computational model and data that offer valuable insights for the refined design of sgRNAs, characterized by heightened sensitivity and specificity. [14].

Operating on a foundation of profound learning principles, the team introduced a groundbreaking framework named DeepCRISPR. This innovation is positioned to predict both the knockout efficacy of sgRNA targets and the profiles of off-target cleavage across the entire genome simultaneously, positioning it in direct competition with established tools. DeepCRISPR leverages deep unsupervised representation learning techniques to autonomously capture the essential sgRNA representation from an expansive collection of genome-wide, unlabeled sgRNAs. Furthermore, they utilize annotated sgRNAs and refine the model through a supervised deep neural network. At the core of DeepCRISPR lies a robust deep learning model, specifically, a deep neural network (DNN). The development of the framework includes autoencoders and convolutional neural networks (CNNs).

Creation of a comprehensive dataset encompassing the knockout efficacy of sgRNA targets across four distinct cell types. This dataset comprises roughly 15,000 sgRNAs validated through experimentation, derived from 1071 genes. Compilation of an exhaustive whole-genome off-target spectrum dataset for human sgRNAs, facilitated by five distinct techniques including GUIDE-seq, BLESS, and IDLV. This compilation includes data from 30 sgRNAs obtained from two distinct cell types.

The model takes as input an extensive dataset of 20-base pair sgRNA sequences that encompass NGG PAM motifs spanning the entirety of the human genome. Each sgRNA sequence is encoded alongside

its corresponding lineage and epigenetic data (Figure 3.a). Moreover, there is an autoencoder to pre-train these unlabeled sgRNA sequences, with the encoder portion as the pre-training network for subsequent stages (Figure 3.c). The fine-tuning of the pre-trained network occurs in tandem with training a novel convolutional network which aids in predicting sgRNA-targeted knockout effects by utilizing labeled datasets containing established knockout efficiencies (Figure 3.b). Researchers further refine two pre-trained networks using miss data sets—one for learning sgRNA sequences and the other for capturing representations of potential miss sites (Figure 3.d). These representations are subsequently concatenated and input into a convolutional network to predict missed areas.

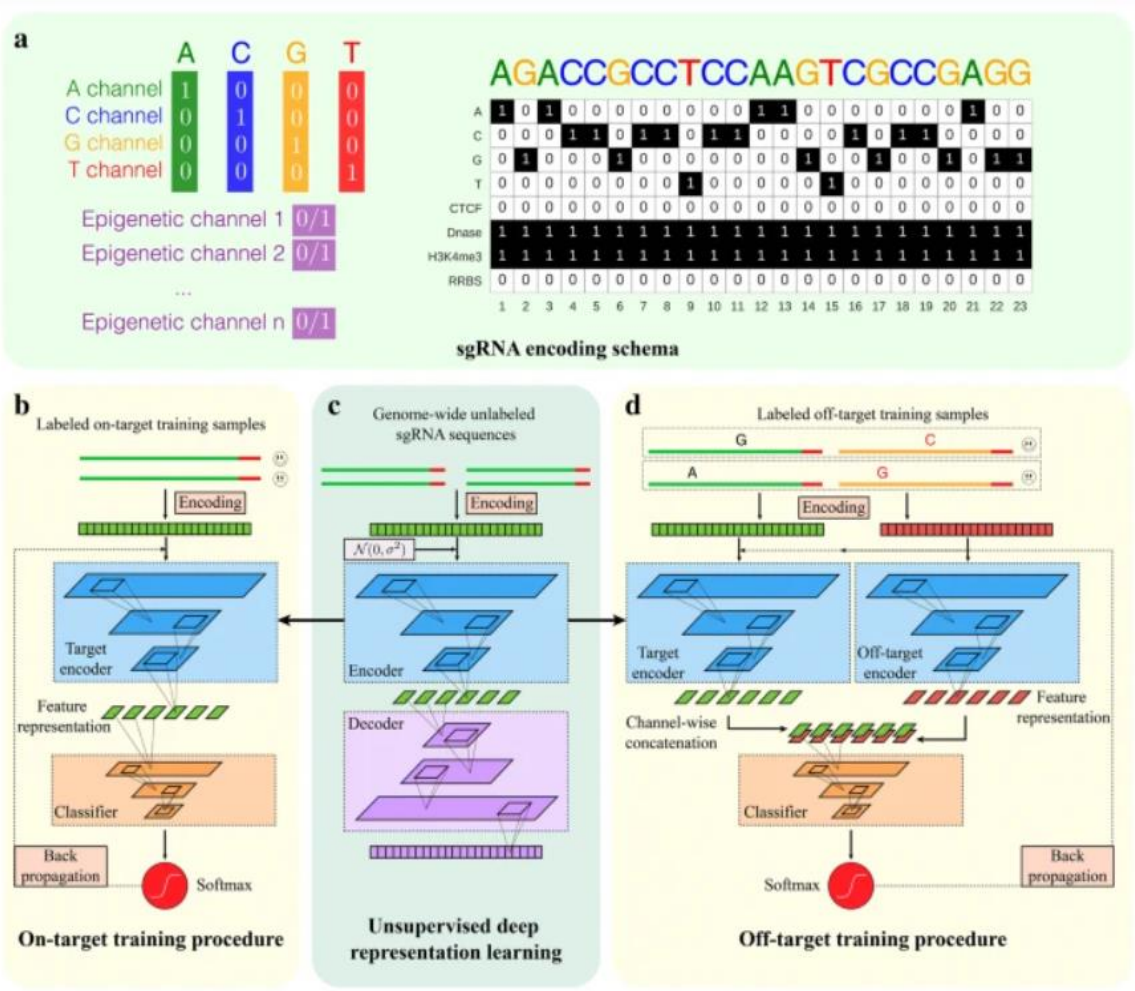

Figure 3. The diagram provides an overview of DeepCRISPR's implementation, encompassing the following aspects; a) Encoding Schema for sgRNA; b) Training Specifications for Predicting sgRNA On-Target Efficacy; c) Self-Guided Deep Representation Learning from Genome-Wide sgRNA Sequences; d) Training Particulars for sgRNA Off-Target Profile Prediction

Employing deep learning models to anticipate sgRNA efficacy showcases superior performance compared to shallow models, particularly when incorporating unsupervised pre-training. DeepCRISPR emerges as a game-changer by obviating the need for manual sgRNA design and excelling in capturing intricate high-level features. Notably, even subtle adjustments to nucleotides distant from the PAM site exert minimal influence on sgRNA potency, indicating a restrained impact on targeting specificity. The significant role of nucleotide preferences is evident in optimizing sgRNA targeting strategies. Furthermore, the alignment of nucleotide preferences with accessible chromatin structures at specific sites presents a promising avenue for advancing the design of sgRNA targets with augmented efficiency and precision.

The versatility of DeepCRISPR in predicting sgRNA efficacy and the profound implications of nucleotide preferences position this novel approach as a transformative tool for advancing the

precision and efficiency of gene editing endeavors. Furthermore, it provides an excellent reference for deep learning compared to other traditional machine learning methods.

3.4. Application of Machine Learning in Predicting Off-Target Effects Case - 4: Leveraging genetic therapeutic effects of CRISPR-Cas9 and Machine Learning for Precise Exon Skipping

A recent MIT study introduces a promising genome editing strategy rooted in CRISPR-Cas9 technology for addressing genetic disorders. This approach harnesses exon skipping and can yield enduring alleviation for hereditary ailments. Scientists leveraged machine learning to cherry-pick Cas9 guide RNAs with the ability to disrupt splice acceptors, thus instigating exon skipping. The detailed experiments on murine embryonic stem cells measured the exon skipping rates across 791 distinct splicing patterns. The results indicate that SkipGuide, the researchers' machine learning framework, aptly identifies efficacious guide RNAs with notable accuracy (0.68 and 0.93 for predicting 50% and 70% thresholds of exon skipping frequencies, respectively). SkipGuide is a machine learning-infused framework tailored to prognosticate the extent of exon skipping prompted by distinct SpCas9 guide RNAs pinpointing splice acceptor sites. SkipGuide amalgamates insights from varied predictive tools like SpliceAI, inDelphi, and MetaSplice, augmenting the precision of forecasts. As a result, SkipGuide holds potential in aiding the identification of appropriate guide RNA candidates to assess therapies involving CRISPR-Cas9-mediated exon skipping [6].

4. Discussion

Despite the powerful effects of machine learning in many of the above aspects, the limitations that remain unresolved make sgRNAs designed by machine learning have flaws that cannot be eradicated. Therefore, summarizing and attempting to address the significant limitations of machine learning in designing efficient sgRNAs will guide future research. In addition to underfitting/overfitting, which is common in other fields of machine learning, as well as the research cost caused by relying on a large amount of prior data [2], this section also summarizes some defects in the scientific research fields discussed in this paper and targeted improvement measures.

Currently, many machine learning methods used to design efficient sgRNA are imbalanced when they come to the dataset; that is, there is a distinct gap between the number of negative and positive samples [8]. The potential explanation could be that the required experimental data published did not contain harmful data or the positive sample needed to be higher due to the low frequency of Homology-Directed Repair [12]. This flaw can lead to bias in the model's prediction of editing efficiency in specific genes or variants and poor model performance due to small sample sizes. This problem may be mitigated using resampling techniques. Balance the number of samples between classes by increasing the number of samples in a few classes or decreasing the number of samples in a majority class. These methods include but are not limited to, random oversampling (increasing Minority samples), random under sampling (decreasing majority samples), and SMOTE (Synthetic minority Over-sampling Technique). At the same time, resampling should also pay attention to the problems such as the imbalance of noise cancellation caused by oversampling. This solution can also be used with efficiency data conversion, converting continuous efficiency data to binary and setting thresholds, using a percentage threshold (such as 50%) to classify the data more evenly. Of course, the interpretability of the model generated by too low a threshold is questionable. Therefore, accumulating wet experiments with rare samples is still necessary, and the threshold-setting standards must be further optimized and improved.

Besides, a unified standard in terms of model, principle, prior data, and data preprocessing is required, which leads to specific differences in the prediction results produced by different machine learning models. For example, given the feature that affects the cutting efficiency of sgRNA, Haeussler et al. [18] found that bumps rarely occurred and the cutting efficiency was negligible. In the data set tested by Abadi et al. [9], the bulge accounted for about 20%. Moreover, most of the raised sgRNAs had

low efficiency in cutting target DNA sequences, and a few raised sgRNAs had medium efficiency. These differences affect the generality of the tool and the reference value of the prediction conclusions and increase the cost of developing more advanced tools for repeated validation. In this scenario, two possible solutions could be: 1), standardized dataset and methodologies are assembled and ready for sharing. This can reduce bias among studies and thus improve the consistency of predictions. Sharing data sets and methods can make researchers' cross-cutting comparisons and validation easier. 2), integrating multiple methods: combining multiple machine learning methods and models can mitigate the impact of differences between different methods. More stable prediction results can be obtained through the integrated method, thus improving the model's reliability.

At present, the off-target effect of sgRNA cannot be entirely or almost eliminated, which means the application of CRISPR gene editing in human clinical trials still has a long way to go, and many characteristics affecting the efficiency of sgRNA also make the research and development of machine learning tend to be differentiated and lax. In this case, borrowing other RNAs used to defend against foreign genetic elements and using machine learning-assisted modification of sgRNAs to weaken their peripheral cutting activity or enhance on-target activity may be another shortcut. As reported from Feng Zhang's team, Fanzor is a guiding endo nuclide for a CRISPR-like defense mechanism in eukaryotes and has no collateral cleavage activity [19]. Therefore, the analysis and comparison of cutting mechanism and activity data of CRISPR-gRNA & Fanzor using machine learning methods may help to modify sgRNA to gain the advantages of other guided RNAs.

5. Conclusion

In conclusion, the fusion of CRISPR-Cas9 gene editing with machine learning offers a transformative pathway to overcome challenges in guide RNA (sgRNA) design. While CRISPR-Cas9 revolutionized genetic manipulation, sgRNA limitations like off-target effects and specificity issues hinder its potential. Machine learning proves a potent remedy, effectively enhancing sgRNA design. Various machine learning models, from gradient boosting to deep learning, have been harnessed to optimize sgRNA efficiency. By leveraging prior data, model construction, and cross-validation, these techniques predict off-target effects, amplify on-target activity, and balance specificity-efficiency trade-offs. Models such as random forests and deep neural networks accurately predict knockout efficiency and exon skipping potential. However, challenges persist, including data imbalance and varying model predictions. Standardizing data methods and integrating models could mitigate these disparities. Additionally, combining CRISPR-gRNA with other RNA guides, like Fanzor, via machine learning-guided modifications offers promising avenues. The synergy of biology and machine learning presents unprecedented possibilities for CRISPR-Cas9 advancements, propelling genetic engineering toward revolutionary biomedical and therapeutic horizons. As interdisciplinary collaboration continues, the harmonization of CRISPR-Cas9 and machine learning has the potential to reshape genetic editing's landscape.

References

[1] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-rna–guided DNA endonuclease in adaptive bacterial immunity. Science, 337 (6096), 816 - 821.

[2] Konstantakos, V., Nentidis, A., Krithara, A., & Paliouras, G. (2022). CRISPR–cas9 gRNA efficiency prediction: An overview of predictive tools and the role of deep learning. Nucleic Acids Research, 50 (7), 3616 - 3637.

[3] Foschi, N., Athanasakis, E., Gasparini, P., Stazio, M. D., & D'Adamo, A. P. (2020). Systematic analysis of factors that improve HDR efficiency in CRISPR / Cas9 technique.

[4] Lisa Li, H., Nakano, T., & Hotta, A. (2013). Genetic correction using engineered nucleases for gene therapy applications. Development, Growth & Differentiation, 56 (1), 63 - 77.

[5] Chen, L., Wang, S., Zhang, Y., Li, J., Xing, Z., Yang, J., Huang, T., & Cai, Y. (2017). Identify key sequence features to improve CRISPR sgRNA efficacy. IEEE Access, 5, 26582 - 26590.

[6] Louie, W., Shen, M. W., Tahiry, Z., Zhang, S., Worstell, D., Cassa, C. A., Sherwood, R. I., & Gifford, D. K. (2021). Machine learning based CRISPR gRNA design for therapeutic exon skipping. PLOS Computational Biology, 17 (1), e1008605.

[7] Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., Lee, S., Yoon, S., & Kim, H. (. (2018). Deep learning improves prediction of CRISPR–cpf1 guide RNA activity. Nature Biotechnology, 36 (3), 239 - 241.

[8] Sherkatghanad, Z., Abdar, M., Charlier, J., & Makarenkov, V. (2023). Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: A review. Briefings in Bioinformatics, 24 (3).

[9] Abadi, S., Yan, W. X., Amar, D., & Mayrose, I. (2017). A machine learning approach for predicting CRISPR-cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLOS Computational Biology, 13 (10), e1005807.

[10] Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-cas9. Nature Biotechnology, 34 (2), 184 - 191.

[11] Koch, B., Nijmeijer, B., Kueblbeck, M., Cai, Y., Walther, N., & Ellenberg, J. (2018). Generation and validation of homozygous fluorescent knock-in cells using CRISPR–cas9 genome editing. Nature Protocols, 13 (6), 1465 - 1487.

[12] Liu, X., Yang, Y., Qiu, Y., Reyad-ul-ferdous, M., Ding, Q., & Wang, Y. (2020). SeqCor: Correct the effect of guide RNA sequences in clustered regularly interspaced short palindromic repeats/Cas9 screening by machine learning algorithm. Journal of Genetics and Genomics, 47 (11), 672 - 680.

[13] Charlier, J., Nadon, R., & Makarenkov, V. (2021). Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-cas9 gene editing. Bioinformatics, 37 (16), 2299 - 2307.

[14] Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B., Gu, F., Qu, S., Huang, D., Wei, J., & Liu, Q. (2018). DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. Genome Biology, 19 (1).

[15] Hatture, S. M., & Kadakol, N. (2021). Clinical diagnostic systems based on machine learning and deep learning. Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics, 159 - 183.

[16] Dhanjal, J. K., Dammalapati, S., Pal, S., & Sundar, D. (2020). Evaluation of off-targets predicted by sgRNA design tools. Genomics, 112 (5), 3609 - 3614.

[17] High-density guide RNA tiling and machine learning for designing CRISPR interference in Synechococcus Sp. PCC 7002. (n.d.).

[18] Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., Joly, J., & Concordet, J. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biology, 17 (1).

[19] Fanzor: First CRISPR-like system found in eukaryotes. (2023). GEN Biotechnology, 2 (4), 276 - 277.