Sequence Analysis

LIGAND: Locus Inference and Generative Adversarial Network for gRNA Design

Woody Hulse [1], Pratham Rathi [1] Taj Gillin [1], Zhen Ren [1]

[1] Brown University, Providence, RI, 02912, USA

Abstract

The advent of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology, notably the CRISPR-Cas9 system—an RNA-guided DNA endonuclease-introduces an exciting era of precise gene editing. Now, a central problem becomes the design of guide RNA (gRNA), the sequence of RNA responsible for locating a bind location in the genome for the CRISPR-Cas9 protein. While existing tools can predict gRNA activity, only experimental or algorithmic methods are used to generate gRNA specific to DNA subsequences. In this study, we propose LIGAND, a model which leverages a generative adversarial network (GAN) and novel attention-based architectures to simultaneously address on- and off-target gRNA activity prediction and gRNA sequence generation. LIGAND's generator produces a plurality of viable, highly precise, and effective gRNA sequences with a novel objective function consideration for off-site activity minimization, while the discriminator maintains state-of-the-art performance in gRNA activity prediction with any DNA and epigenomic prior. We also apply pertubation analysis to understand and validate model performance. This dual functionality positions LIGAND as a versatile tool with applications spanning medicine and research.

Code available at `https://github.com/woody-hulse/LIGAND`.

1 Introduction

The discovery of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology has transformed the field of genomics, providing a tool for precise gene editing and manipulation. While there have been significant recent advances in the technology (Hillary & Ceasar 2020), the primary protein used is CRISPR-Cas9, an RNA-guided DNA endonuclease. Initially discovered as a bacterial defense mechanism in the adaptive immune system of archaea and bacteria, the editing capabilities have proved revolutionary. The CRISPR system has been used to develop treatments and cures for various illnesses—cancer, metabolic disorders, infectious diseases, and genetic diseases—and advance agricultural development (Nidhi et al., 2021).

The protein itself is directed to specific genomic sequences by a guide RNA (gRNA). The gRNA sequence is critical to determining the specificity of the CRISPR-Cas9 system, as it determines the ultimate binding site of the protien. The sequence is designed as a complement to the target DNA sequence, and a correct pairing leads to a proper double-strand break at the intended location. Once broken, the cell's repair machinery allows for subsequent modifications.

While the CRISPR-Cas9 system is powerful, the design of gRNA is a difficult, non-trivial task. One must design a sequence that is both specific and effective, taking into account factors such as sequence composition, chromatin structure, GC-content, secondary structure formation, and, critically, potential off-site effects (Mohr 2016). Previous approaches use either experimental methods (Zhang et al., 2015) to manually examine or algorithmic methods to predict these effects. These methods are either tedious or lack proper gRNA/DNA distribution coverage to be applied in real-time experimental settings. Therefore, there is a strong need for generalizable models that can synthesize gRNA and evaluate efficacy for given DNA sequences that yield both high specificity and generalizeability.

There are several unexplored innovations to the problem approach that we address with our problem, LIGAND. We create generative and predictive models which leverage novel deep learning structures such as attention and hierarchical convolution. These show promise for being able to translationally relate sequential information better than simple dense, RNN, or convolutional approaches. Reliance on techniques such as convolution (Dipankar et al., 2022) and LSTM (Niu et al., 2023) only may have trouble relating subsequences over long ranges or analyzing more sophisticated structure within sequential data. Further, a generative network hasn't yet been adversarily trained; rather, a simple ground truth is used to measure viability. This doesn't reflect the nature of the problem: for any particular DNA site, there will likely be several viable gRNA that show high activity. Thus, there is a need for a generative model that can creatively identify multiple potential gRNA with consideration of on- and off-target efficacies.

2 Related Work

2.1 Relevant models

We identify two widely-cited models in the field that address similar problems of predicting on- and off-site gRNA activity.

DeepHF (Daqi et al., 2019) is a design tool for specific Cas9 nucleases, eSpCas9(1.1), SpCas9-(HF1) and wild-type SpCas9. For initial gRNA activity prediction, the study one-hot encodes the sequence, which is then embedded to create a new representation. These embeddings are then passed into a bi-directional LSTM. The new encoding is then concatenated with hand-crafted biological features and finally serves as an input to a fully connection layer that creates a prediction score. The team used these

scores to curate an augmented dataset of gRNA sequences for these three Cas9 nucleases.

DeepCRISPR (Guohi et al., 2018) is a convolution-based deep-learning online service developed for gRNA design. It uses epigenetic information as well as base sequence information to predict gRNA activity across a particular region. The first step of the process takes in .68 billion sgRNA sequences that are associated with epigenetic information gathered across different cell types. The unlabeled sequences are then used for pretraining an autoencoder to learn denoised encodings. These encodings are then used as input for a convolutional neural network. The entire hybrid neural network is then trained on a much smaller set of labeled data, thus learning optimal weights for the CNN and also fine-tuning weights for the autoencoder parent network. The resultant hybrid neural network can then create predictions for on-target site activity and the parent network was later reused with another CNN to do the same for off-target site activity.

3 Methods

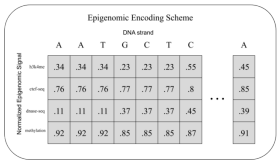
3.1 Data collection and preprocessing

We accumulate genomic and epigenomic data from the GRCh38 genome representation from the NIH. We use this genome representation for backward compatibility with in-class models and for its higher precision at the nucleotide scale. Further, because gRNA binding is dependent on factors such as histones, transcription factors, DNase, and other parts of the epigenome, those are included in the generative input and discriminative prior as well.

We selected CTCF-seq and H3K4me3 histone modification from Chip-Seq data, chromatin opening information from DNase-seq data, and DNA methylation from RBBS sequencing data. The data was selected from the GM12878 cell line, all drawn from ENCODE and NIH human genome databases. We binarize epigenomic signal values due to their sparsity throughout the genome, thus choosing to give them more importance when present. Further, noised pertubations on the input would otherwise likely greatly impact the model's interpretation of the epigenomic context.

Figure 1. gRNA encodings

Figure 2. Epigenomic encodings



There are several data sources containing experimental gRNA activity on certain parts of the genome (Pulecio et al., 2018). In particular, we use the dbGuide database (Gooden et. al., 2021), a vast selection of experimentally-validated guide RNAs for associated DNA sequences. This database was selected in particular because of its breadth over the distribution of nucleotide combinations for potential gRNA synthesis. This is important, as it allows our trained model to gain a better representation of a full distribution, limiting out-of-distribution issues and fully unlocking the breadth of our generative approach.

3.2 Model design

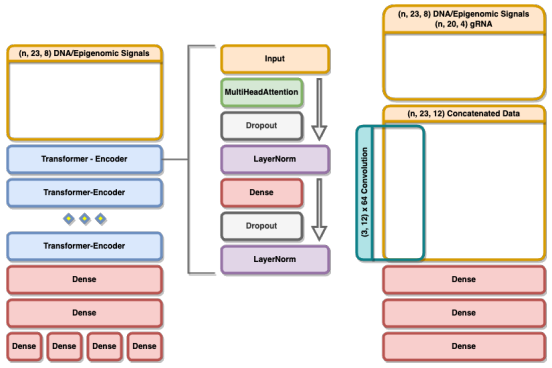
A generative-adversarial network (GAN) is a model architecture which employs two distinct models–a generator and a discriminator, to play a two-player game for the generation of some data. The generator is tasked with creating output to which the discriminator assigns a score of 1, while the discriminator tries to assign scores of 1 to ground truth outputs and assign 0 to generated inputs. This competing scheme is captured in the objective function, where the loss function for each respective model is adversarially oriented against the other.

LIGAND modifies the formulation of generative-adversarial structure similar to a cGAN (conditional-GAN), where both the generator and discriminator are provided with a DNA/epigenomic prior. The nature of the predicted gRNA output (probability distributions over base classes) presents unique challenges for framing a reasonable objective function, and loss function modifications are key in capturing each use-case for LIGAND.

The generator in LIGAND is tasked to generate gRNA given a DNA and epigenomic signal input. The generator relies on a novel attention-based architecture, leveraging the attention mechanism to find meaningful relationships in DNA and epigenetic signal input. Our motivation for using an attention mechanism here is to comprehend both the structures in DNA and the latent contributors in gRNA, such as secondary structures and GC-content, which contribute to the efficacy and soundness of the gRNA design. The head of both models uses dense layers to generate a prediction (and a final time distributed dense layer for the generator).

Figure 3. Generator architecture

Figure 4. Discriminator architecture



Both models consume the $(n, 23, 8)$ array of one-hot encoded DNA and signal epigenomic signals, which both use as a prior for their respective prediction tasks. The generator outputs an $(n, 20, 4)$ gRNA prediction, where each column of the ith matrix corresponds to a predicted probability distribution for the base at that position. For LIGAND's discriminator, convolution is applied over the concatenated and aligned DNA/epigenomic/gRNA pairing to identify kernels which can properly synthesize their compatibility. We then interpret the probability output of the discriminator as an activity score for a particular gRNA on a given site in the genome.

The DNA prior plays a pivotal role in LIGAND's discriminator model. As we will demonstrate, this DNA prior allows the discriminator to generalize to entirely new segments of the DNA, allowing virtually any DNA/gRNA pair to be tested for efficacy.

3.3 Training

For each network, we compute the loss as follows for any given experimentally validated pair of DNA (X_{DNA}) and gRNA (X_{gRNA}) or non-pair ($X_{\text{DNA [rand]}}$):

$$L_{\text{BCE}} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log(1 - \hat{Y}_i))$$

$$L_{\text{CE}} = -\sum_{i=1}^{n}(Y_i \cdot \log \hat{Y}_i)$$

$$\begin{aligned}
L_G &= \lambda_{G_1} L_{\text{BCE}}(1, D(G(X_{\text{DNA}}), X_{\text{DNA}})) && \text{Activity of predicted gRNA} \\
&+ \lambda_{G_2} L_{\text{CE}}(X_{\text{gRNA}}, G(X_{\text{DNA}})) && \text{Similarity to experimental gRNA} \\
&+ \lambda_{G_3} L_{\text{BCE}}(0, D(G(X_{\text{DNA}}), X_{\text{DNA [rand]}})) && \text{Discourage offsite activity} \\
L_D &= \lambda_{D_1} L_{\text{BCE}}(1, D(X_{\text{gRNA}}, X_{\text{DNA}}) && \text{Activity of correct gRNA} \\
&+ \lambda_{D_2} L_{\text{BCE}}(0, D(G(X_{\text{DNA}}), X_{\text{DNA}}) && \text{Catch generated gRNA} \\
&+ \lambda_{D_3} L_{\text{BCE}}(0, D(X_{\text{gRNA}}, X_{\text{DNA [rand]}})) && \text{No activity with random bps}
\end{aligned}$$

The interplay of loss functions establishes form a complex relationship between the two sub-models of LIGAND. On one hand, like any GAN, the generator tries to trick the discriminator while the discriminator tries to catch the generator. On the other hand, the generator importantly relies on the discriminator to guide its gRNA generation to minimize offsite activity. These separate our model from other algorithmic methods, while still being motivated by accuracy.

To train LIGAND, we begin with the DNA/epigenomic signal pairing. Before passing this data into the generator, we apply a noise function $X_{\text{DNA}} = X_{\text{DNA}} + \gamma\mathcal{N}(0, 1)$ where γ is a hyperparameter ($\gamma = 0.1$). This perturbation of the model inputs allows the generator to exercise more creativity and improve robustness in the prediction of gRNA.

Next, we pass this data into the discriminator. Because our generator outputs a probability vector with discrete ground truth values, we assign X_{gRNA} the properly ordered probability values of the generated output to ensure we remain in distribution for the discriminator comparison. Otherwise, comparing discrete ground truth values with distributions would be trivial for the discriminator model. We then compute the loss terms and backpropagate across both models.

4 Experiments and Results

4.1 Selection of model architecture

In order to evaluate the viability for different model architectures to complete this task, we propose three formulations for each: a baseline MLP approach, a convolution (and deconvolution) approach, and a transformer approach. For comparison, we isolate each model to evaluate it on generalized train and test data.

Figure 5. Generator validation loss

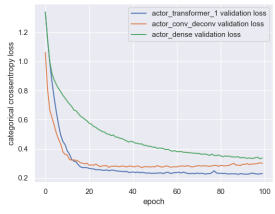
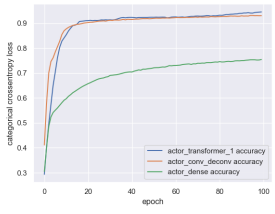


Figure 6. Generator validation accuracies



For the generator model, we ask each model to generate a gRNA sequence from a given DNA/epigenomic location, comparing categorical crossentropy loss against the ground truth gRNA and evaluating accuracy based on the highest probability category. The MLP used is a simple, 3 layer network with 64 dense units in each hidden layer. The convolution/deconvolution network contains three convolution layers, a latent dense layer of size 76, and three more inverse convolution layers. The attention encoder architecture is identical to the generator architecture above. The experimental results were fairly conclusive, showing that the transformer formulation confirms our hypothesis that this particular problem benefits from the existence of an attention map.

Table 1. Metrics after 50 epochs across various architectures.

Generator	Discriminator	AUROC	Generator Accuracy	Discriminator Accuracy
MLP	MLP	0.5041	0.225	0.318
ConvDeconv	CNN	0.9762	0.888	**0.718**
Transformer	CNN	**0.9878**	**0.927**	0.546
Transformer	Transformer	0.9867	**0.927**	0.673

Figure 7. Discriminator loss

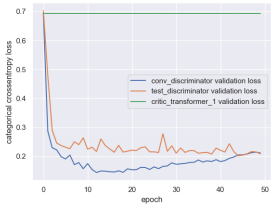
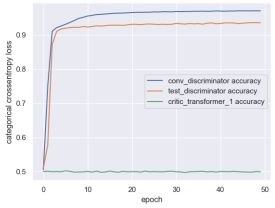


Figure 8. Discriminator accuracies



To test potential discriminator architectures, we consider the set of paired DNA/gRNA labeled with 1 as well as a set of perturbed gRNA (3 base pair perturbations) with label 0. Our architectures are virtually unchanged other than the end dense size of $23 \rightarrow 1$ and removing the deconvolution layers from the convolution model (as detailed in Methods). For this problem, the convolution discriminator shows the best performance. It should be noted that the transformer is still a work in progress, but as several constructions of it also failed to learn meaningful discriminative relationships and the training process is about 5x longer, we exclude it from consideration for now.

4.2 GAN accuracy validation

We next pair our attention encoder generator with the convolution discriminator to create LIGAND, the combined model. We want to first show that the accuracy for both models, though following a roughly adversarial path (particularly the discriminator), is high. Accuracy for the generator is the average value of the highest categorical base output against the true base of the gRNA:

$$A_G = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}(\text{argmax}Y_{ik} = \text{argmax}\hat{Y}_{ik})$$

For the discriminator, we want to identify true gRNA while catching generated gRNA:

$$A_D = \frac{1}{2n}\sum_{i=1}^{n}(1 - \text{round}D(\hat{Y}_i)) + \frac{1}{2n}\sum_{i=1}^{n}(\text{round}D(Y_i))$$

As shown in Table 1, the transformer generator with a deconvolution discriminator achieves the highest accuracy for the generator and the highest Area Under the ROC Curve (AUROC) score. As outlined in the methods section, our goal is for the discriminator to predict both gRNA on-site activity and off-site activity. A higher AUROC score indicates greater sensitivity to potential off-target effects (See Table 1).

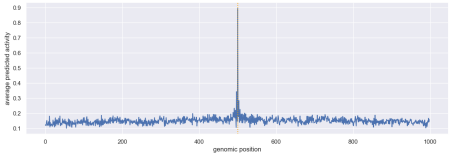
4.3 Evaluating on- and off-site efficacy

Next, we evaluate the performance of the discriminator in a more practical setting. The original goal of the discriminator is to identify both on- and off-target effects of DNA/gRNA pairings. We can evaluate this by treating the discriminator as a sliding window over the DNA such that the resultant discriminant value is the assessed probability of a given gRNA activity at a particular site.

$$\text{Activity}_i = D(X_{\text{gRNA}}, X_{\text{DNA}_i})$$

Plotting these activity scores for some examples, we can visualize that the average activity score on test gRNA sequences is correctly located at the bind site for multiple scales, as desired, typically with very low probability of off-site effects.

Figure 11. Average predicted activity (1000bp)



We can further use this method to look at specific gRNA examples to analyze both the efficacy of the gRNA at the desired bind site as well as potential off-target binding sites.

Figure 12. Validated examples

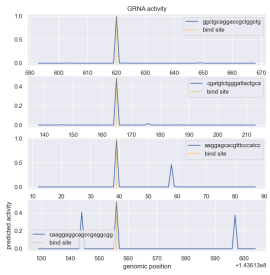
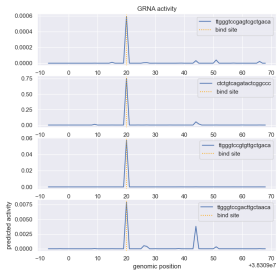


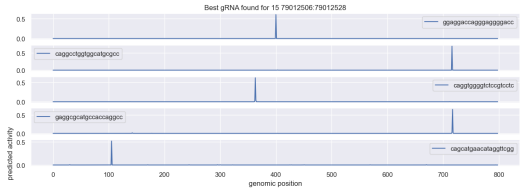
Figure 13. Generated examples



For gRNA design, we can apply LIGAND's generator. By applying different noise samples $\mathcal{N}(0, 1)$ to a given gRNA input, we can generate multiple candidate gRNA from the generator. Then, by applying the discriminator, LIGAND provides a way to compare this gRNA for both its on-target and off-target effects. In Figure 14, four example gRNA are compared, where the highest activity and lowest off-target activities are observed.

In practical applications, we often design gRNA to target specific genes, particularly for gene knockout tasks. This involves cleaving in specific regions while maintaining minimal off-site effects in other regions. We can perform a search over a desired window for generated gRNA and rank-order them with on/off-site efficacy using the discriminator. Our model successfully outputs high-accuracy predictions with virtually no off-site effects over large genomic regions.

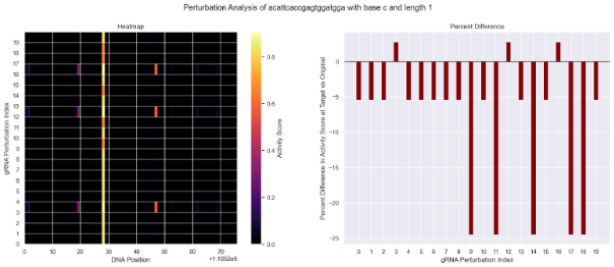
Figure 14. Top-5 generated examples for example gene region



4.4 Perturbation analysis

We can also perturb input to evaluate the differential effects of gRNA design on predicted efficacy. By holding a target DNA site constant, we took a known, matched gRNA strand and perturbed it on a base-by-base resolution, allowing us to pass the perturbed gRNA and target DNA to the discriminator and analyze differences in activity scores. The images below show the activity graph across the DNA sequences for two original gRNA strands and then the effects of perturbing them with the different bases, of which only a few are shown. The heatmaps show that perturbations still maintain the bind site as the region of highest activity on the DNA strand. Some other details to be noticed are that the activity score at the target is more sensitive to perturbations with the bases "T" and "C", as that is where we see the largest decreases in activity. Furthermore, the different strands react differently to the perturbations, suggesting the importance of each base is likely dependent on the larger context of the gRNA strand.

Figure 15. Example perturbation map



5 Discussion

In this study we introduce LIGAND, a generative adversarial network that addresses both gRNA activity prediction and gRNA generation in concert. We determined a hybrid attention-based and convolutional approach to generating and discriminating gRNA through experimentation with subtasks, a result reinforced by the improved performance of this union over other GAN formulations. We affirm discriminator accuracy by observing the highest activity scores correctly concentrated around the binding site of the DNA sequence. Additionally, perturbation analysis enhanced the interpretability of LIGAND, highlighting our model's robust ability to identify influential gRNA variants.

As a demonstration of the practical application of LIGAND, we demonstrate that noised generator priors can effectively stimulate a plurality of output predictions, creating a multitude potential gRNA candidates. From these, we are able to analyze activity with the discriminator across the genome. These examples show that this quickly training associated generative model could aid experimental researchers in significantly narrowing down the search space for viable gRNA sequences, including finding entirely novel sequences.

Looking ahead, we aim to complete a transformer discriminator, although the incomplete version already outperformed the deconvolution model in preliminary experimentation. We haven't explored experimentally verified off-target activity scores between mismatched gRNA and DNA sequences in our model due to lack of experimentally validated datapoints (relative to the scale of our gRNA reference database), anticipating that it will provide additional insights into the model's abilities and interpretation.

We envision that this structure could extend beyond gRNA research to other DNA editing endeavors, such as applying a similar model to investigate gene expression prediction. By enabling unlimited and effective gRNA design and identifying potential side effects, we hope LIGAND can help to advance genome editing methods for research and clinical applications.

6 Contributions

Woody Hulse	Generator/discriminator architecture designs, objective function engineering, DNA + gRNA preprocessing, activity experiments, drafting
Pratham Rathi	Epigenomic preprocessing, interpretability/perturbation analysis, figures
Taj Gillin	DeepCRISPR port, pertubation analysis, DNA + gRNA preprocessing, code/architecture design
Zhen Ren	Analysis, drafting, literature review

7 References

1. Mohr SE, Hu Y, Ewen-Campen B, Housden BE, Viswanatha R, Perrimon N. CRISPR guide RNA design for research applications. FEBS J. 2016 Sep;283(17):3232-8. doi: 10.1111/febs.13777. Epub 2016 Jun 22. PMID: 27276584; PMCID: PMC5014588.

2. Nidhi S, Anand U, Oleksak P, Tripathi P, Lal JA, Thomas G, Kuca K, Tripathi V. Novel CRISPR-Cas Systems: An Updated Review of the Current Achievements, Applications, and Future Research Perspectives. Int J Mol Sci. 2021 Mar 24;22(7):3327. doi: 10.3390/ijms22073327. PMID: 33805113; PMCID: PMC8036902.

3. Hillary, V.E., Ceasar, S.A. A Review on the Mechanism and Applications of CRISPR/Cas9/Cas12/Cas13/Cas14 Proteins Utilized for Genome Engineering. Mol Biotechnol 65, 311–325 (2023). https://doi.org/10.1007/s12033-022-00567-0

4. Mohr SE, Hu Y, Ewen-Campen B, Housden BE, Viswanatha R, Perrimon N. CRISPR guide RNA design for research applications. FEBS J. 2016 Sep;283(17):3232-8. doi: 10.1111/febs.13777. Epub 2016 Jun 22. PMID: 27276584; PMCID: PMC5014588.

5. Nidhi S, Anand U, Oleksak P, Tripathi P, Lal JA, Thomas G, Kuca K, Tripathi V. Novel CRISPR-Cas Systems: An Updated Review of the Current Achievements, Applications, and Future Research Perspectives. Int J Mol Sci. 2021 Mar 24;22(7):3327. doi: 10.3390/ijms22073327. PMID: 33805113; PMCID: PMC8036902.

6. Hari Shrawgi, et al. "Convolution Neural Network Model for Predicting Single Guide RNA Efficiency in CRISPR/Cas9 System." *Chemometrics and Intelligent Laboratory Systems*, Elsevier, 25 Apr. 2019, www.sciencedirect.com/science/article/abs/pii/S0169743918306117.

7. Zubaida Sa'id Ameen a, et al. "C-SVR CRISPR: Prediction of CRISPR/CAS12 Guiderna Activity Using Deep Learning Models." *Alexandria Engineering Journal*, Elsevier, 27 Feb. 2021, www.sciencedirect.com/science/article/pii/S1110016821000788.

8. Baisya, Dipankar, et al. "Genome-Wide Functional Screens Enable the Prediction of High Activity CRISPR-Cas9 and -CAS12A Guides in Yarrowia Lipolytica." *Nature News*, Nature Publishing Group, 17 Feb. 2022, www.nature.com/articles/s41467-022-28540-0.

9. Chuai, Guohui, et al. "DEEPCRISPR: OPTIMIZED CRISPR Guide RNA Design by Deep Learning - Genome Biology." *BioMed Central*, BioMed Central, 26 June 2018, genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1459-4.

10. Kim, Hui Kown, et al. "SpCas9 Activity Prediction by Deepspcas9, a Deep Learning ... - Science." *Science.Org*, 6 Nov. 2019, www.science.org/doi/10.1126/sciadv.aax9249.

11. Konstantakos, Vaileios, et al. "CRISPR–Cas9 gRNA Efficiency Prediction: An Overview of Predictive Tools and the Role of Deep Learning." *Academic.Oup.Com*, 22 Apr. 2022, academic.oup.com/nar/article/50/7/3616/6555429.

12. Luo, Jiesi, et al. "Prediction of Activity and Specificity of CRISPR-Cpf1 Using Convolutional Deep Learning Neural Networks - BMC Bioinformatics." *BioMed Central*, BioMed Central, 13 June 2019, bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2939-6.

13. Obeid, Ahmad, and Hasan Al-Marzouqi. "CRISPR-Vae: A Method for Explaining CRISPR/CAS12A Predictions, and an Efficiency-Aware Grna Sequence Generator." *bioRxiv*, Cold Spring Harbor Laboratory, 1 Jan. 2021, www.biorxiv.org/content/10.1101/2021.07.05.451176v1.full.

14. Wang, Daqi, et al. "Optimized CRISPR Guide RNA Design for Two High-Fidelity Cas9 Variants by Deep Learning." *Nature News*, Nature Publishing Group, 19 Sept. 2019, www.nature.com/articles/s41467-019-12281-8.

15. X;, Niu R;Peng J;Zhang Z;Shang. "R-CRISPR: A Deep Learning Network to Predict off-Target Activities with Mismatch, Insertion and Deletion in CRISPR-Cas9 System." *Genes*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/34946828/. Accessed 13 Oct. 2023.

16. Pulecio J, Verma N, Mejía-Ramírez E, Huangfu D, Raya A. CRISPR/Cas9-Based Engineering of the Epigenome. Cell Stem Cell. 2017 Oct 5;21(4):431-447. doi: 10.1016/j.stem.2017.09.006. PMID: 28985525; PMCID: PMC6205890.

17. Alexander A Gooden, Christine N Evans, Timothy P Sheets, Michelle E Clapp, Raj Chari, dbGuide: a database of functionally validated guide RNAs for genome editing in human and mouse cells, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D871–D876, https://doi.org/10.1093/nar/gkaa848