Guide Sequence Attention & Biological Context

Improve Evaluation of CRISPR Off-Target Cleavage

WillB

Will Bednarz, Paul Krupski, Jonah Schwam, Jonathan Dou

Brown University | CSCI 2952G | Deep Learning in Genomics

https://github.com/pjkrupski/CS2951G_ClassProject.git

Introduction

The field of genome engineering transformed in 2014 with the advent of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) gene editing systems [1]. CRISPR systems have emerged as potent tools for genome engineering, enabling cost-effective, high-accuracy, site-specific modifications of genes [2]. The versatility of CRISPR technology has cemented its place in biomedical research [3].

Researchers utilizing CRISPR systems program the system to target and modify a specific DNA sequence of interest. The CRISPR system is programmed by synthesizing a 20-25 bp guide RNA (gRNA) that binds to the CRISPR-associated protein (cas) [4]. Ideally, cas will then bind to the DNA region complementary to the gRNA and induce the intended gene modification; however, CRISPR is imperfect and may inadvertently modify other regions of the genome that bear similarity to the target sequence. Given that CRISPR is a gene-modifying system, these off-target events result in unintended and potentially hazardous mutations, thus compromising the safety and feasibility of CRISPR-based interventions [5].

Researchers have developed tools to accurately predict off-target sites to bolster the reliability and safety of CRISPR systems. Existing tools for off-target prediction primarily rely on heuristic algorithms and rule-based systems. For instance, industry-standard tools, namely Cas-OFFinder and the CFD scoring algorithm, were developed to provide a predictive analysis of potential off-target sites based on sequence similarity and known CRISPR-Cas9 behavior. However, these tools often require extensive computational resources and may not capture the complex underlying biological interactions that contribute to off-target activities [6, 7, 8].

Another promising avenue to tackle the challenge of off-target prediction is the employment of deep learning methodologies, which have exhibited remarkable success in various domains of bioinformatics, including other applications of the CRISPR system. Notably, piCRISPR, R-CRISPR, and DeepCRISPR are existing deep-learning tools that aid researchers in implementing CRISPR gene editing systems more effectively [9, 10, 11].

DeepCRISPR is engineered to optimize the design of gRNAs in CRISPR for both on and off-target efficacy. It comprises three sub-models: a Sequence Model for identifying local sequence patterns in gRNAs, an Epigenetic Model that incorporates epigenetic information from ChIP-seq and ATAC-seq data, and an Image Model that processes Hi-C images of local genomic structure to capture spatial relationships among genomic elements. PiCRISPR, a CNN-RNN-based model, is designed to predict unintended genetic modifications, including off-target events, that may arise from CRISPR modifications. R-CRISPR is a similar deep-learning-based model aimed at predicting unintended genetic modifications from CRISPR. The piCRISPR model encodes potential gRNAs and target sequences as binary matrices and employs a CNN for feature extraction, followed by an RNN for off-target activity predictions. For our proposed model, WillB, we built and expanded upon the capabilities of current off-target prediction models.

Related Work

CRISPR-Net.

CRISPR-Net is a deep-learning-based model to quantify CRISPR off-target activity. As an input, the model takes in two separate one-hot encoded sequences. The first is the guide RNA sequence that would be used to program a CRISPR system, and the second is the target DNA sequence in a cell's genome that the CRISPR system is designed to target. The model outputs a single off-target classification score for gRNA-target pairs based on sequence mismatches and indels. CRISPR-Net focuses on two different types of mismatches: alternative sequences with a missing base, and alternative sequences with an added base. CRISPR-NET was validated with the CIRCLE-Seq and the SITE-Seq database, which are included in the data used to train and validate the WillB model.

CRISPR-Net utilizes a CNN and a bidirectional LSTM. CRISPR-Net's use of an LSTM raises the notion that long-term genomic dependencies may contribute to the off-target activity of a CRISPR system. Figure 1 illustrates the architecture for CRISPR-Net.

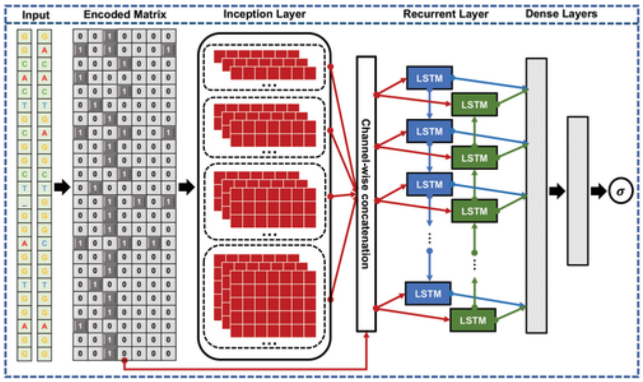

Figure 1. CRISPR-Net Architecture features a CNN followed by a bidirectional LSTM and dense layers for off-target predictions.

R-CRISPR.

R-CRISPR similarly models the risk of off-target CRISPR activity by classifying gRNA-target pairs as high or low risk. As is done with CRISPR-Net, R-CRISPR categorizes off-target activities into sequences with a base mismatch from the target site, sequences with a base deleted from the target site, and sequences with a base added from the target site. R-CRISPR was trained on CIRCLE, PKD, PDH, SITE, and GUIDE_I datasets before validation on GUIDE_II and GUIDE_III datasets. Similar to CRISPR-Net, R-CRISPR utilizes a CNN layer for feature extraction followed by an RNN for generating predictions (Figure 2).

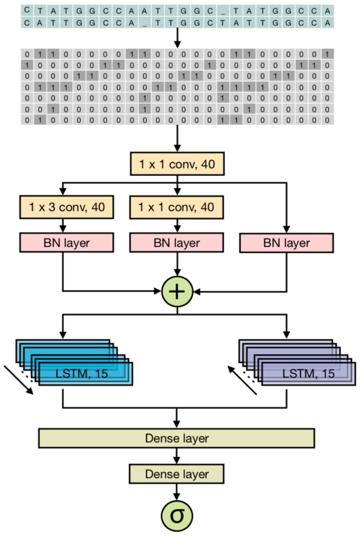

Figure 2. R-CRISPR Architecture features a CNN for feature extraction followed by a bidirectional LSTM and dense layer for off-target predictions.

In validation, CRISPR-Net and R-CRISPR both exhibit poor AUPRC performance, indicating a weak ability to classify positive examples. In the following research, we identify potential sources of error present in CRISPR-Net and R-CRISPR. We identify a model, piCRISPR, that rectifies some of the shortcomings of these models, and we then proceed to expand upon piCRISPR to further improve its performance.

3. Methods

We introduce three significant additions to existing models, including CRISPR-Net and R-CRISPR, to improve their predictive performance. First, the existing models primarily utilize target sequences to generate predictions. By incorporating additional biologically relevant features into the model input, WillB will be able to more effectively capture the complex, multifaceted nature of CRISPR activity. Namely, piCRISPR and WillB include nucleosomal position scores, which quantify the target gene sequence's physical orientation, and CRISPROff free energy scores, which quantify the chemical binding affinity between the guide RNA sequence and the target DNA sequence. Second, the low AUPRC score in existing models may be attributed to class imbalance during training. By bootstrapping our dataset for training, we may be able to improve WillB's performance on positive examples. Lastly, CRISPR activity is often dominated by specific subsets or bases within the target sequence. By incorporating a transformer into WillB, WillB will be able to focus on the regions of the guide sequence that dominate CRISPR system behavior, thereby improving the model's predictive accuracy. A transformer will still capture the long-range dependencies captured by the LSTMs of CRISPR-Net and R-CRISPR with the additional benefit of an attention mechanism to capture the dominant regions of the guide sequence.

Baseline Model.

To implement the new features of WillB, we iterated upon an existing model, piCRISPR. As mentioned previously, piCRISPR is a model designed to combine the hybrid CNN-LSTM architecture of the previous models with a more diverse input dataset. The baseline implementation of WillB is grounded in the piCRISPR model [10]. The piCRISPR implementation includes various baseline models and model iterations for comparison. Their most complex model is a hybrid RNN-CNN model, which is the highest-performing model for this prediction task to date. In addition to incorporating the additional input parameters of piCRISPR and following similar data preprocessing steps as piCRISPR, WillB will utilize a transformer architecture.

Model Inputs & Outputs.

Preceding models, including R-CRISPR and CRISPR-Net, exclusively utilize a one-hot encoded guide sequence and target sequence as input. piCRISPR and WillB expand the input feature space to include nucleosomal positioning of the target sequence (BDM, NuPoP, GC147) and the free energy of the guide-target binding (CRISPROff estimated free energy). The physical orientation of the target sequence and the chemical binding affinity of the guide sequence to the target dominate the CRISPR system's binding efficacy and provide critical biological context. The output of R-CRISPR, CRISPR-Net, piCRISPR, and WillB are all single neurons for binary classification.

The authors of the piCRISPR paper generate training labels by binarizing off-target activity. They threshold cleavage activity (CA) below the lowest reported assay accuracy of 10^{-5}, creating two labels for the training data: 0 for low cleavage activity, and 1 for high cleavage activity [10].

Replicating piCRISPR & Data Preprocessing.

First, the training process for two of the simpler, RNN-based models if piCRISPR was replicated. The two chosen models were two-layer and three-layer networks, with training and validation results included in Table 1. The models were trained for 100 epochs, within which their validation accuracy plateaued. Of note, the validation accuracies were higher than anticipated, indicating overfitting in the training process. Further examination of the dataset indicated a class imbalance and lack of data shuffling in the original piCRISPR data preprocessing. Negative samples represent 92% of the original dataset. As such, an additional balanceClasses function was implemented to increase the representation of off-site examples using bootstrapping methods.

This finding represents an additional significant improvement of WillB over piCRISPR: the balanceClasses function helps mitigate the class imbalance of piCRISPR's training data, preventing overfitting and potentially improving WillB's performance on positive samples. The balance Classes function works by extending the training data to contain putative off-target sites within a certain threshold for a given number of mismatches. The custom dataset contains approximately 35,000 samples.

Table 1. 2-layer and 3-layer RNN performance indicates overfitting due to data class imbalance.

	2-Layer RNN	**3-Layer RNN**
Train Loss	0.0272	0.0064
Accuracy (%)	91.56	96.79

Transformer Implementation.

After reimplementing the baseline model outlined in piCRISPR and adjusting their data processing, the next step was to incorporate a transformer-based model that would match the input and output shape of piCRISPR.

The adjusted WillB architecture features two new linear layers, one to serve as an embedding to reshape its 139-dimensional input to a vector of size 512 for the transformer layer, and a second to map from the transformer output to the classification output neurons of the model. Like CRISPR-Net, R-CRISPR, and piCRISPR, WillB's output neuron will classify the risk of off-target activity for the input guide RNA and target DNA pair.

The transformer was designed to mimic the encoder-only design of BERT. Since the model input features now include nucleosomal positioning and CRISPROff free energy, the input features are no longer sequential. Thus, WillB does not include the positional encoding layer of BERT. WillB's encoder stack contains two encoder layers and 4 attention heads for each layer (Figure 3).

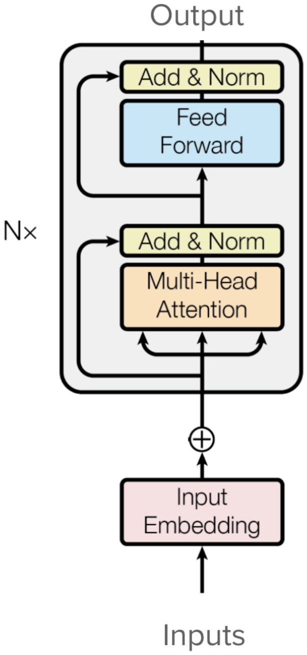

Figure 3. WillB mimics the architecture of BERT, excluding the positional encoding layer.

Results

Our preliminary results underscore the role of the underlying data in enhancing CRISPR off-target prediction accuracy. The WillB model, built on the piCRISPR foundation, demonstrates a promising classification accuracy of 95.5%. Thus, WillB outperforms several other models including the 2-layer RNN, 3-layer RNN, CNN, and hybrid CNN-RNN outlined in piCRISPR (Figure 4). To date, the piCRISPR hybrid CNN-RNN remains the state-of-the-art model for CRISPR off-target prediction. Thus, further validation of

WillB’s performance could support its claim to be the dominant model for this classification task. Although WillB outperforms piCRISPR by only 0.3% in classification accuracy, this difference is substantial in the context of precision and safety.

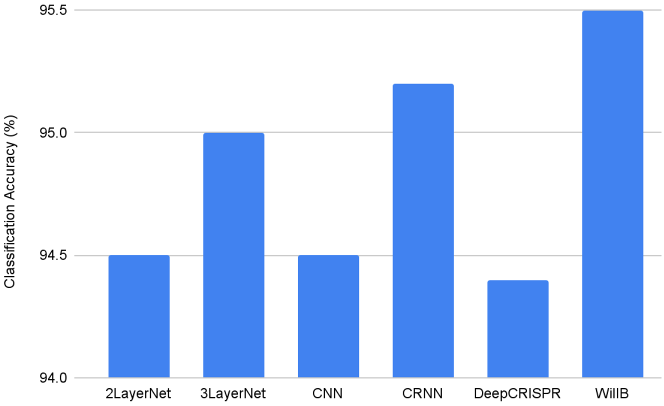

Figure 4. WillB comparative performance in classification accuracy.

Discussion

The emergence of CRISPR as a transformative tool in the biomedical sciences underscores the need for accurate and reliable predictive models to assess the risks of CRISPR-based interventions. WillB, our novel deep-learning model, attempts to bridge the gap left by existing tools like R-CRISPR and CRISPR-Net. It incorporates additional biological data, addresses the class imbalance in training data, and utilizes a transformer architecture to focus on the dominant regions of the guide sequence, addressing several of the shortcomings of existing models.

WillB's slight edge in classification accuracy over its predecessors highlights the critical importance of biological context for bioinformatic models like WillB. Of interest, when optimizing WillB, we found that a 4-layer transformer outperformed more expansive architectures, indicating that the data and classification task could easily overfit. This finding is critical for the development of future models that seek to improve the classification accuracy of CRISPR systems.

Future Directions.

Before considering improvements or other models for this task, further validation of WillB is required. Namely, we intend to perform an ablation study to determine the contribution of each of the unique changes to WillB in improving the model’s performance. In other words, we aim to determine which of our data preprocessing, additional biological context, and transformer architecture is primarily responsible for the observed performance improvement. To that end, we intend to perform AUPRC analysis to quantify how much our modified data preprocessing improved the classification for positive samples. Similarly, multiple rounds of training and validation would support the statistical soundness of WillB’s claimed performance improvement.

Following further validation of the model, we intend to visualize the attention of WillB. First, we can validate that WillB is paying attention to the regions of the guide sequence that dominate CRISPR behavior with biological experiments. Once validated, visualizing the attention mechanism in WillB can provide valuable biological insights into which parts of the guide RNA sequence or genomic context the model deems most significant. This could lead to a better understanding of CRISPR dynamics and guide the design of more effective gRNAs in biology research.

Future iterations of WillB could benefit from including more comprehensive biological data, such as Hi-C data or DNA methylation states. These two data types are of particular interest, as both could inform the physical accessibility of target genomic regions. These factors are increasingly recognized as important in CRISPR-Cas9 targeting. Lastly, while currently focused on CRISPR-Cas9 data. Expanding WillB to accommodate other CRISPR systems, like CRISPR-Cpf1, could broaden its utility across various research and therapeutic contexts.

Conclusion.

The initial results from WillB are promising, indicating a potential step forward in the predictive modeling of CRISPR off-target effects. By integrating improved data processing, additional biological context, and a transformer architecture, WillB represents a novel approach to the computational risk analysis of CRISPR systems. The development of such models is critical for the development of safer and more effective genome editing applications in medicine and biology research.

References

[1] Wang, Joy Y., and Jennifer A. Doudna. “CRISPR technology: A Decade of genome editing is only the beginning.” Science, vol. 379, no. 6629, 2023, https://doi.org/10.1126/science.add8643.

[2] Singh, Vijai. “An introduction to genome editing CRISPR-Cas Systems.” Genome Engineering via CRISPR-Cas9 System, 2020, pp. 1–13, https://doi.org/10.1016/b978-0-12-818140-9.00001-5.

[3] Lanese, Nicoletta, and Aparna Vidyasagar. “What Is CRISPR, the Powerful Genome-Editing Tool?” LiveScience, Purch, 13 Mar. 2023, www.livescience.com/58790-crispr-explained.html#:~:text=CRISPR%20is%20a%20powerful%20tool,potential%20applications%2C%20including%20correcting.

[4] Nidhi, Sweta, et al. “Novel CRISPR–Cas Systems: An updated review of the current achievements, applications, and Future Research Perspectives.” International Journal of Molecular Sciences, vol. 22, no. 7, 2021, p. 3327, https://doi.org/10.3390/ijms22073327.

[5] Coelho, M.A., De Braekeleer, E., Firth, M. et al. CRISPR GUARD protects off-target sites from Cas9 nuclease activity using short guide RNAs. Nat Commun 11, 4132 (2020). https://doi.org/10.1038/s41467-020-17952-5

[6] Listgarten, J., Weinstein, M., Kleinstiver, B.P. et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat Biomed Eng 2, 38–47 (2018). https://doi.org/10.1038/s41551-017-0178-6

[7] Jiecong Lin, Ka-Chun Wong, Off-target predictions in CRISPR-Cas9 gene editing using deep learning, Bioinformatics, Volume 34, Issue 17, September 2018, Pages i656–i663, https://doi.org/10.1093/bioinformatics/bty554

[8] Bao, X.R., Pan, Y., Lee, C.M. et al. Tools for experimental and computational analyses of off-target editing by programmable nucleases. Nat Protoc 16, 10–26 (2021). https://doi.org/10.1038/s41596-020-00431-y

[9] Chuai, G., Ma, H., Yan, J. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol 19, 80 (2018). https://doi.org/10.1186/s13059-018-1459-4

[10] Störtz, Florian, et al. “PICRISPR: Physically informed Deep Learning Models for CRISPR/Cas9 off-target cleavage prediction.” Artificial Intelligence in the Life Sciences, vol. 3, 2023, p. 100075, https://doi.org/10.1016/j.ailsci.2023.100075.

[11] Niu, R.; Peng, J.; Zhang, Z.; Shang, X. R-CRISPR: A Deep Learning Network to Predict Off-Target Activities with Mismatch, Insertion and Deletion in CRISPR-Cas9 System. Genes 2021, 12, 1878. https://doi.org/10.3390/genes12121878

[12] Gligorijević, V., Renfrew, P.D., Kosciolek, T. et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 12, 3168 (2021). https://doi.org/10.1038/s41467-021-23303-9

[13] Florian Störtz, Jeffrey K. Mak, Peter Minary, piCRISPR: Physically informed deep learning models for CRISPR/Cas9 off-target cleavage prediction, Artificial Intelligence in the Life Sciences, Volume 3, 2023, 100075, ISSN 2667-3185, https://doi.org/10.1016/j.ailsci.2023.100075

[14] Niu, R.; Peng, J.; Zhang, Z.; Shang, X. R-CRISPR: A Deep Learning Network to Predict Off-Target Activities with Mismatch, Insertion and Deletion in CRISPR-Cas9 System. Genes 2021, 12, 1878. https://doi.org/10.3390/genes12121878

[15] Chuai, G., Ma, H., Yan, J. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol 19, 80 (2018). https://doi.org/10.1186/s13059-018-1459-4

[16] Florian Störtz, Peter Minary, crisprSQL: a novel database platform for CRISPR/Cas off-target cleavage assays, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D855–D861, https://doi.org/10.1093/nar/gkaa885