A Transformer-Based Approach with Focal Loss for Improved Off-Target Detection in CRISPR-Cas9 Systems

Ishaan Nirmal[1]

Received January 26, 2026
Accepted July 5, 2026
Electronic access August 15, 2026

Genetic disorders affect over 350 million people worldwide, yet effective curative treatments remain limited. CRISPR/Cas9 gene editing offers a promising pathway to correct harmful mutations, but its clinical safety is constrained by off-target edits, which are unintended genomic cuts that can lead to serious biological consequences. Thus, these off-target edits need to be monitored for and detected. However, current off-target detection models are hindered by severe class imbalance between rare true off-target sites and abundant negatives, limited architectural capacity to capture long-range sequence dependencies, and overfitting to specific experimental assay distributions, resulting in poor generalization and low accuracy on true off-target events. This project aims to develop an improved, machine-learning-based off-target detection model that surpasses existing published models in accuracy, balance, and efficiency by using balanced data, a unique base model, and parameter and focal-loss experimentation. Using the rebalanced CCLMoff benchmark dataset, the data were divided into subsets corresponding to different detection methods (GUIDE-seq, DIS-seq/DISplus-seq, and DIG-seq). The rebalancing allowed for better model results, but reduced the direct comparability between this model and published ones. Models were trained using a leave-one-method-out strategy to evaluate generalizability across techniques. The model used was a deep learning model based on DistilBERT. The regular model achieved an accuracy of 85.95% and a PRC AUC of 0.6157, substantially outperforming most published models, particularly in predicting true off-targets. This score was 0.3367 better than the best model available. These results demonstrate that the system forms the basis for a future end-to-end sgRNA design platform.

Introduction

Motivation

Currently, there are many diseases that are caused by genetic factors, such as Cystic Fibrosis, Sickle Cell Anemia, and Duchenne's Muscular Dystrophy[1]. These conditions affect millions worldwide and can be addressed by the advent of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)[2] gene-editing technology[3]. CRISPR directly addresses the root cause of the issue by editing genomes. CRISPR uses a guide RNA and the enzyme Cas9 to cleave DNA at a specific target site. The cell's natural repair mechanisms then fix the break, allowing scientists to inactivate a gene, correct a mutation, or insert new genetic material[4]. The major flaw in this technology, however, is off-target editing[5]. Off-target edits are when CRISPR makes cuts in unintended locations. Due to the genome's sensitivity, even minor edits can lead to significant side effects, including mutations, oncogene activation, immune responses, and/or gene activation or deactivation[6]. Such edits can be as deadly as the genetic disorders themselves. Computational tools have therefore been developed to identify potential off-target sites and support safer clinical use of CRISPR, but the complexity and scale of genomic sequence data make accurate prediction infeasible without machine learning. Machine learning is employed because it is significantly more efficient and can analyze large-scale datasets at a level that is impractical for manual human analysis. However, current machine learning models for detecting off-target edits in CRISPR are inadequate due to inconsistent and imbalanced performance[7]. This is further explored in the next section.

Background

A substantial body of prior work has sought to address CRISPR off-target detection[8]. This involves sequence-based prediction where one guide sequence and one target sequence are provided. The model then must determine whether the pair of sequences is an off-target pair. Early computational tools focused on rule-based or heuristic approaches[9]. Cas-OFFinder[10] performs exhaustive genome-wide searches to identify potential off-target sites. While this approach achieves high recall, it yields an overwhelming number of false positives at higher mismatch thresholds — for example, Cas-OFFinder identified 294,534 candidate sites across just nine guide RNAs at a 6-mismatch threshold, of which only

[1] *Plano West Senior High School, Texas, USA*

354 were validated as true off-target sites, a false positive rate exceeding 99.8%[11], where a mismatch refers to a nucleotide difference between the guide RNA and the genomic DNA sequence, and cannot model bulges, chromatin effects, or the positional importance of mismatches, resulting in near-zero precision. CRISPOR[12] and related scoring approaches, including the MIT score and CFD score[13], were trained on limited experimental datasets that are biased toward GUIDE-seq[14] data. These models perform poorly on off-target sites with many nucleotide differences between the guide RNA and the DNA, and their shallow architectures oversimplify the complex dynamics of Cas9–DNA binding. COSMID[15] generates extremely large candidate sets without probabilistic scoring of cleavage likelihood. As a result, it cannot meaningfully rank candidate sites by biological relevance. Later deep learning approaches, such as DeepCRISPR[16], introduced a hybrid framework combining a Deep Convolutional Denoising Neural Network autoencoder with a CNN to jointly predict on-target knockout efficacy and off-target cleavage from sgRNA–DNA sequence pairs. However, because CNN-based architectures extract features through local convolutional filters, they are limited in their ability to model long-range dependencies between positions across the full guide–target alignment — an important signal, since off-target activity is influenced by interactions between mismatch positions distributed across the entire 20-nucleotide guide sequence, not only locally adjacent ones. These models also exhibit strong dependence on training assay composition, poor interpretability, and reduced performance when evaluated on datasets from different laboratories. DeepSpCas9[17] and DeepHF[18] further improved sequence-based prediction but were primarily optimized for on-target activity rather than off-target detection. Their off-target performance degrades with increasing mismatches between nucleotides in the sequence pair, and they do not incorporate chromatin or epigenetic context. CRISPR-Net[19] introduced hybrid architectures that combine multiple types of neural network layers, such as CNNs for local sequence feature extraction and RNNs or Transformers for modeling long-range dependencies, but it requires strict input formatting and complex, highly specific data. Attention-based models, such as AttnToMismatch[20], can overfit to dataset-specific mismatch distributions and require large, balanced datasets that are often unavailable in practice. More recently, CCLMoff[21] proposed a transfer-learning approach, but it was trained on datasets in which negative samples far outnumbered positive ones. In the leave-one-dataset-out evaluation on the GUIDE-seq dataset, the same protocol used in this work, CCLMoff achieved a reported AUPRC of 0.513 and an overall F1-score of 0.429, indicating limited performance on the minority positive class. Results for several leading models in the field are shown in Figure 1. This shows AUPRC and Receiver Operating Characteristic (ROC) performance on the dataset using GUIDE-seq data with the Leave-One-Out strategy. Across all prior approaches, two structural problems recur. First, severe class imbalance, with positive-to-negative ratios documented as high as 1:4189, causes models trained under standard conditions to be biased toward the majority class, achieving high overall accuracy while performing poorly on the minority class of true off-target events, as consistently reflected in low F1-scores and AUPRC values across the models[3]. Second, CNN-based architectures cannot efficiently model long-range dependencies across the guide–target sequence pair, limiting their ability to capture distributed mismatch patterns that span the full alignment. The present model directly addresses both problems: the class imbalance problem through training set rebalancing to a 1:5 positive-to-negative ratio combined with focal loss and class weighting, and the architectural problem through DistilBERT[22], a transformer encoder whose self-attention mechanism models pairwise dependencies across all sequence positions simultaneously.

Objective

While transformer- and attention-based architectures have previously been applied to CRISPR off-target prediction[21], existing models either rely on computationally expensive domain-specific encoders or do not systematically combine long-range sequence modeling with explicit imbalance-correction strategies. This study tests the hypothesis that a computationally efficient distilled transformer encoder (DistilBERT[22]), when combined with training set rebalancing, focal loss[24], and class weighting, achieves higher AUPRC and macro F1-score than CCLMoff[21] under an identical leave-one-dataset-out evaluation on the GUIDE-seq held-out dataset. This hypothesis is falsifiable: if the proposed model does not exceed CCLMoff's reported AUPRC of 0.513 and F1-score of 0.429 under the same evaluation conditions, it is not supported. The most crucial problem in current models is their poor performance on the minority (true off-target edit) class, as evidenced by low F1- Scores and low Area Under the Precision-Recall Curve (AUPRC). In this case, this implies a high false-negative rate. A high false-negative rate indicates that a model fails to detect off-target edits in regions where they occur. Such faulty detection can lead to unnecessary medical operations, which can be as dangerous as the initial genetic disorder itself. Thus, the specific goal of the project is to develop a model with higher Macro F1 and AUPRC scores. Such a model will achieve more balanced predictions by correctly identifying both positive and negative samples, outperforming CCLMoff[21], the current state-of-the-art model evaluated under the same protocol. This will be achieved by using a more balanced dataset, a unique base model that uses a pretrained distilled transformer to capture long-range guide–target sequence dependencies and better model distributed mismatches that CNN-based approaches

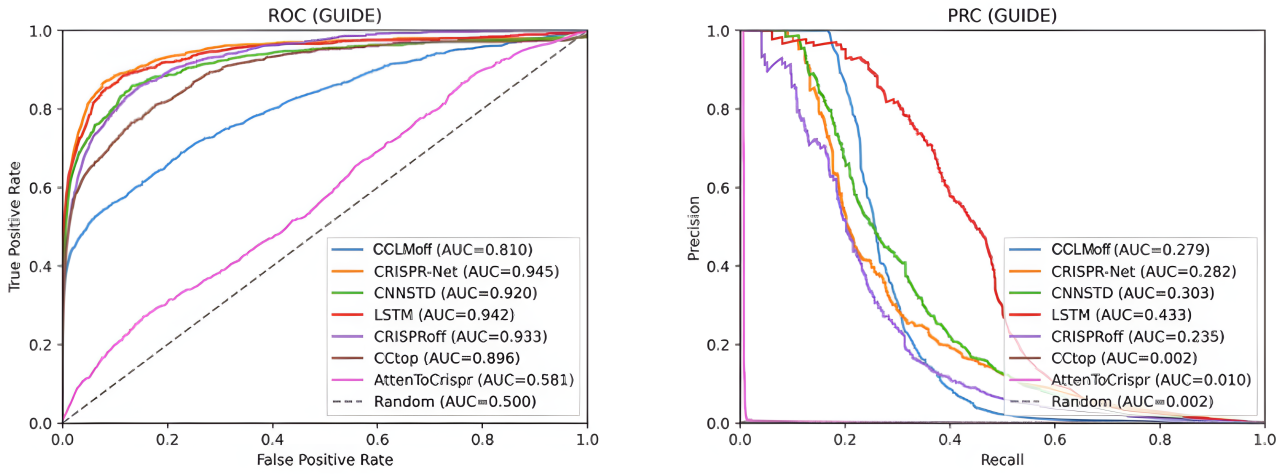

Figure. 1 Best models currently available and their results on the Leave One Dataset Out method using the GUIDE-seq dataset for evaluation [21].

struggle to represent, and by experimenting with techniques and parameters such as focal loss and class weights.

Methods

Data

The data used for this project comes from a comprehensive dataset [21] that includes several genome-editing methods. This is the dataset that the CCLMoff paper, the most recent major model in the field, assembled and used for testing. It comprises of sgRNA–target DNA sequence pairs derived from 13 experimental off-target detection assays across 418 guide RNAs. Positive samples (label = 1) correspond to experimentally validated off-target cleavage events confirmed by sequencing read evidence in the original assay; negative samples (label = 0) correspond to candidate genomic sites identified by Cas-OFFinder that were not detected as active in the experimental assay. It contains 6,393,373 samples and 11 features per sample, such as chromosome, location, orientation, and length. There are 6,354,054 negative (not true off-target) samples as compared to only 39,319 positive (true off-target) samples. The PAM types represented are NGG (SpCas9). The dataset includes 11 methods for detecting where CRISPR makes cuts and off-target edits. These methods are experimental sequencing assays that identify CRISPR-induced DNA cleavage events using different strategies, such as tagging double-strand breaks in living cells or sequencing cleavage products generated in vitro. Because each assay differs in sensitivity and biological realism, they produce distinct off-target profiles. The counts for each of these methods are shown in Table 1 below.

Table 1 The sample counts for each sequencing method

Sequencing Method	Sample Count
CHANGE-seq	1,376,264
SURRO-seq	1,013,017
GUIDE-seq	520,281
SITE-seq	61,221
BLESS	45,477
IDLV	43,441
DIS-seq	37,564
HGTS	12,820
Digenome-seq	307
DISplus-seq	45
Extru-seq	1

Data Preprocessing

Of the 11 features, two were used for model training: the guide and target sequences. These were concatenated into input_seq with ’ — ’ as a separator between the sgRNA and target sequences. This character was chosen because it does not appear in nucleotide sequences, ensuring it is unambiguously interpreted by the tokenizer as a boundary signal between the two sequences. For model training, the dataset was rebalanced by randomly sampling five negative samples for each positive sample. This created a more balanced dataset, which likely would lead to better results for the F1-score and AUPRC. This ratio was set to 5:1 to ensure that class imbalance was not severe while preserving data variation. A 1:1 ratio maximizes minority-class signal but departs most dramatically from the native class distribution and risks overfitting to a small posi-

tive set. Ratios closer to the native distribution (1:161) render positive-class learning nearly infeasible, as demonstrated by the near-zero recall reported for models trained without rebalancing. The 5:1 ratio preserves meaningful negative-class variation. The proposed model and CCLMoff[21] were trained under different sampling conditions, as this work applies 1:5 dataset-level rebalancing while CCLMoff uses batch-level bootstrapping on the native distribution. However, both models are evaluated on the same held-out GUIDE-seq test set at native class prevalence, without any rebalancing applied at evaluation time. The performance comparison therefore reflects differences in minority-class detection on an identical, unmodified test set, which is a methodologically valid basis for comparison despite the difference in training-stage sampling strategies. The DNA sequence, which is the target sequence, is then tokenized at the character level using the AutoTokenizer from the DistilBERT base model[22], and every instance of 'T' (Thymine) is replaced with 'U' (Uracil). The rationale for this transformation is as follows. The sgRNA sequence in the input is an RNA molecule and naturally contains U rather than T. By converting T to U in the target DNA sequence as well, both the guide and target subsequences are expressed in the same four-character alphabet A, C, G, U, creating a uniform input representation and eliminating a spurious asymmetry between the two halves of the concatenated input. Under DistilBERT's character-level tokenization scheme, 'T' and 'U' are treated as two distinct tokens drawn from the model's general English vocabulary and neither carry biological pretraining knowledge, so the conversion does not conflict with pretrained embeddings and instead ensures that identical nucleotide positions in the guide and target are represented by the same token type. All input sequences were tokenized at the character level by inserting a space between each character prior to passing the sequence to the DistilBERT AutoTokenizer. Tokenized inputs were padded to a maximum length of 128 tokens using the tokenizer's [PAD] token, and sequences exceeding this length were truncated from the right. Given that each sgRNA and target sequence is 23 nucleotides, the spaced character representation of the full concatenated input (sgRNA + separator + target) produces a token sequence of approximately 97 characters, well within the 128-token limit, meaning truncation was not applied to any sample in practice. Ambiguous nucleotides represented by the character 'N' in the dataset were retained as-is and treated as a distinct character token, allowing the model to learn any positional associations with ambiguity rather than discarding such positions.

Model

Architecture

The model employs an encoder–classifier architecture, which is well-suited to sequence-based prediction tasks such as CRISPR off-target edit detection. This approach follows a transfer-learning framework, leveraging a pretrained encoder to extract rich, contextualized representations from raw nucleotide sequences, which are then mapped to a binary off-target prediction by a task-specific classifier. Separating representation learning from classification allows the model to capture complex sequence relationships without overfitting the final decision layer, which is especially important given the limited number of positive off-target examples. Using a pretrained transformer encoder is particularly advantageous for this project because transformers can model long-range dependencies across the entire sequence. Off-target activity is determined not solely by local mismatches but also by interactions across the entire guide–target alignment. The self-attention mechanism in the DistilBERT base model[22] enables the model to weigh the relative importance of different positions in the sequence, thereby directly addressing limitations of earlier CNN-based approaches that relied on local receptive fields.

Backbone Model

DistilBERT[22] was selected as the backbone model because it provides a strong balance between representational power and computational efficiency. As a distilled transformer, it retains much of the performance of larger language models while reducing parameter count and training cost. This is well aligned with the project's goals, which emphasize robustness and generalization over excessive model complexity. Rather than applying a pretrained pooling layer or averaging token embeddings, the model directly uses the CLS token embedding as the sequence representation. The CLS token is designed to encode global information across the input sequence, making it well-suited for sequence-level classification. Bypassing the pretrained pooler avoids introducing architectural assumptions learned from natural language tasks that may not transfer to biological sequence data, allowing the classification head to learn task-specific decision boundaries from a more general sequence representation. A potential concern with using a transformer pretrained on English corpora is the mismatch between natural language and nucleotide sequences. However, under the character-level tokenization scheme employed here, the model receives individual nucleotide characters as tokens, not linguistic subwords, meaning the pretrained vocabulary and token embeddings are repurposed as a general-purpose lookup table rather than relied upon for linguistic knowledge. The primary capability being transferred from pretraining is not semantic understanding of English text but rather the self-attention mechanism's learned ability to model pairwise contextual dependencies across token positions, which is architecture-level rather than content-level knowledge. This rationale is consistent with prior work showing that general-purpose transformer architec-

Table 2 A comprehensive summary of the model architecture and training configuration is provided

Category	Property	Value
Backbone	Model	DistilBERT-base-uncased
	Number of transformer layers	6
	Number of attention heads	12
	Hidden size (embedding dim)	768
	Intermediate (FFN) size	3072
	Total parameters	˜66M
	Pretrained parameters	Fine-tuned (not frozen)
Input	Maximum sequence length	128 tokens
	Tokenization	Character-level with spaces
	Sequence representation	CLS token embedding
	Pretrained pooler	Bypassed
Classifier Head	Dropout rate	0.2
	Output	Binary (off-target / not off-target)
Training	Optimizer	AdamW
	Learning rate	210^{-5}
	Weight decay	0.01
	Batch size	32
	Maximum epochs	50
	Early stopping patience	3 epochs (on validation loss)
	Gradient clipping	1.0
	LR scheduler	StepLR (step=2, γ=0.5)
	Random seed	42
Loss	Loss function	Focal loss
	Focal α	0.75
	Focal γ	2.0
	Class weighting	Inverse class frequency

tures, even without domain-specific pretraining, can achieve strong performance on genomic sequence classification tasks. The CCLMoff paper itself demonstrated that its 'CCLMoff-Vanilla' variant, trained from scratch on off-target data without RNA-FM pretraining, still achieved considerable performance, which the authors attributed to the strength of the transformer framework itself[21].

Training

The model was trained using a leave-one-out method. This is where the model is trained on a dataset that excludes one method and then tested on a dataset containing data for that method. The GUIDE-seq method was excluded because results were publicly available for other models that employed the same strategy. This was done to make the comparison between this model and others clearer and more objective. The training dataset consisted of 229,176 samples and included data from all methods except GUIDE-seq. There were 190,980 negative samples and 38,196 positive samples, consistent with a 5:1 ratio. The dataset was further split: 90% of the data was used solely for training, and the remaining 10% for validation. The test dataset comprised only data obtained using the GUIDE-seq method. This dataset contained 6,738 samples, of which 5,615 were negative, and 1,123 were positive, also at a 5:1 ratio.

To address the severe class imbalance inherent in CRISPR off-target datasets, we employ focal loss[23], defined as:

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

where p_t denotes the predicted probability of the true class, α_t is a class-balancing factor that adjusts the relative contribution of positive and negative samples, and γ is a focusing parameter that down-weights well-classified examples.

Focal loss mitigates extreme class imbalance by down-weighting easy, correctly classified negative samples and focusing learning on hard-to-classify positive samples (true off-target sites), enabling the model to learn more effectively under conditions where negatives vastly outnumber positives. Additionally, the positive class was up-weighted by the inverse of its class frequency, thereby increasing the contribution of rare off-target events to the loss. Together, these mechanisms encourage the model to prioritize difficult positive ex-

amples and improve recall while maintaining a balanced precision–recall trade-off. This choice aligns directly with the project's objective of improving F1-score and AUPRC, which are sensitive to performance on the minority class. The batch size of 32 represents a balance between computational efficiency and stable gradient estimation under GPU memory constraints. The AdamW optimizer[24] was selected for its effectiveness in training transformer-based architectures and for its use of decoupled weight decay, which helps prevent overfitting. A StepLR scheduler[25] was applied to decay the learning rate at fixed intervals, allowing for larger updates early in training and finer adjustments as convergence is approached. Training was conducted for up to 50 epochs, with early stopping based on validation loss, using a patience of three epochs. This strategy prevents overfitting and ensures that the selected model generalizes well to unseen datasets. The best-performing model on the validation set was saved and used for final evaluation on the held-out test data.

Results

Model performance was evaluated on the held-out GUIDE-seq test set, which comprised 6,738 samples at a 5:1 negative-to-positive ratio (5,615 negative and 1,123 positive samples). All reported metrics include 95% confidence intervals derived from 1,000 stratified bootstrap resamples.

The model achieved an AUPRC of 0.6157 (95% CI: 0.5877–0.6421) and a positive-class F1-score of 0.5701 (95% CI: 0.5447–0.5928), with precision of 0.5815 (95% CI: 0.5530–0.6101) and recall of 0.5592 (95% CI: 0.5291–0.5866) on the true off-target class. The ROC-AUC was 0.8632 (95% CI: 0.8526–0.8739). The negative-class F1-score was 0.916, the macro-average F1-score was 0.7431, and the weighted-average F1-score was 0.8584. Full per-class metrics are shown in Table 2, the precision-recall curve in Figure 2, and the ROC curve in Figure 3.

These figures compare favourably against prior models evaluated under the same leave-one-dataset-out protocol on GUIDE-seq. MOFF achieved an AUPRC of 0.282, CRISPR-IP 0.337, CRISPR-DNT 0.381, and CCLMoff 0.513 [20]. The proposed model's AUPRC of 0.6157 is the highest reported under this protocol.

In the context of genome-wide CRISPR screening, these results carry a specific practical meaning. Alignment tools such as Cas-OFFinder can identify hundreds of thousands of candidate off-target sites per guide RNA at a six-mismatch threshold, the vast majority of which are false positives. A computational model operating in this setting functions primarily as a ranking and prioritisation tool, not a binary arbiter, since experimental validation of all candidates is infeasible. From this perspective, the positive-class recall of 0.5592 indicates that the model recovers approximately 56% of true off-target sites in the test set, and the AUPRC of 0.6157 reflects consistent precision at moderate recall thresholds. The clinical significance of the false negative rate is asymmetric: a missed true off-target site in a coding region, proto-oncogene, or tumour suppressor poses a greater safety risk than a false positive, which simply adds a site to the experimental validation queue[22]. The improvement in positive-class F1-score over CCLMoff therefore represents a concrete reduction in the rate of missed true off-target events, which is the error type of greatest concern in therapeutic applications of CRISPR. This interpretation is directly supported by the CNN ablation results. Despite being trained under identical conditions, the 1D CNN baseline achieved an AUPRC of 0.1319 and a positive-class F1-score of 0.0054, collapsing to near-universal positive prediction with a precision of 0.0027. This confirms that the imbalance-correction strategies alone are insufficient without an architecture capable of representing the distributed, long-range mismatch patterns that distinguish true off-target events from negative candidates, and that the performance gains of the proposed model are attributable to the transformer architecture rather than the training strategy. The model is intended as an improved computational screening step that more reliably flags candidates for downstream experimental confirmation, not as a replacement for such confirmation.

Class	Precision	Recall	F-1 Score	Support
Negative (0)	0.9125	0.9195	0.9160	5615
Positive (1)	0.5815	0.5592	0.5701	1123
Accuracy			0.8595	6738
Macro Average	0.747	0.7394	0.7431	6738
Weighted Average	0.8573	0.8595	0.8584	6738

Figure. 2 Classification Metric Report for GUIDE-seq

Similarly, the process was carried out using DIS-seq and DISplus-seq as the methods instead of GUIDE-seq. The results on this method from the paper's model and the state of the art models are shown below.

1D CNN Baseline Comparison

To isolate the contribution of the transformer architecture from the imbalance-correction training strategy, a 1D CNN baseline was trained under conditions identical to the proposed model, including the same 1:5 rebalanced training set, focal loss (α=0.75, γ=2.0), class weighting, AdamW optimizer, and held-out GUIDE-seq test set. The CNN baseline achieved an

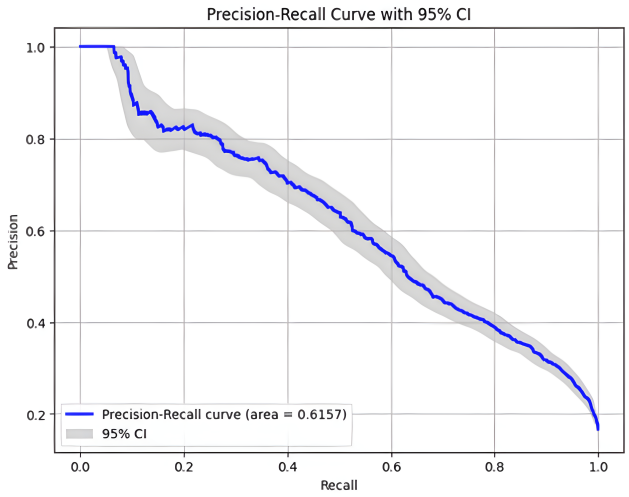

Figure. 3 Precision-Recall Curve for GUIDE-seq

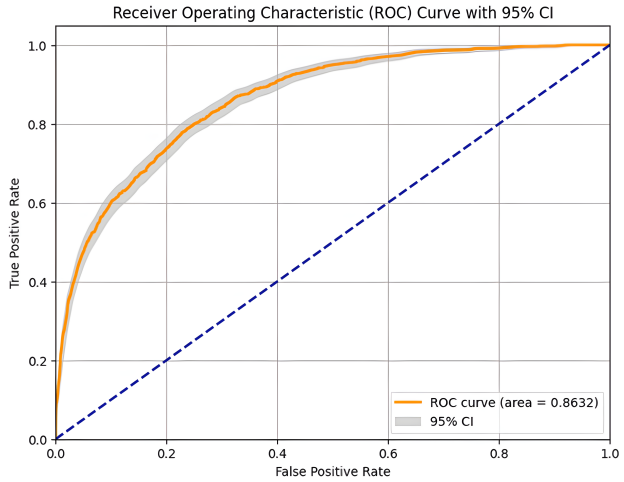

Figure. 4 Receiver Operating Characteristic Curve for GUIDE-seq. The dashed line indicates random chance.

AUPRC of 0.1319, a positive-class F1-score of 0.0054, precision of 0.0027, and recall of 0.8994, with a ROC-AUC of 0.7466. The near-zero precision and F1-score indicate that the CNN defaulted to predicting the majority of samples as positive, demonstrating an inability to learn a discriminative decision boundary despite the imbalance-correction strategies applied during training. By contrast, the proposed DistilBERT model achieved an AUPRC of 0.6157 and a positive-class F1-score of 0.5701 under identical training conditions. The magnitude of this gap (0.4838 AUPRC and 0.5647 F1) demonstrates that the performance of the proposed model is attributable to the transformer architecture's capacity to model long-range dependencies across the full guide-target sequence pair, and not to the rebalancing or loss function alone.

Discussion

For GUIDE-seq, The AUPRC for this model was 0.6157, which is higher than that of any model presented in Figure 1, including CCLMoff, the state-of-the-art model in the field. The F1-score on the negative (majority) class is 0.916, while the F1-score on the positive (minority) class is 0.5701. The weighted average F1-score for this model is 0.8584, while the macro average F1-score is 0.7431. Similarly, for DIS-seq and DISplus-seq, the AUPRC was 0.8357, significantly higher than the best current model, which was 0.681. CCLMoff reported an AUPRC of 0.513 and a positive-class F1-score of 0.429 under the same evaluation protocol[21]. The proposed model improves on both figures, achieving an AUPRC of 0.6157 and a positive-class F1-score of 0.5701. Similarly, for DIS-seq and DISplus-seq, the AUPRC was 0.8357, significantly higher than the best current model, which was 0.681. However, an AUPRC of 0.6157 and 0.8357 are moderate in absolute terms. Precision declines at higher recall thresholds, which means the model is best understood as a prioritisation and ranking tool for downstream experimental validation rather than a standalone binary classifier at native genome-wide prevalence.

When interpreting these results, it is important to consider the effect of the negative-to-positive sample ratio used during training and evaluation. A perfectly balanced 1:1 ratio would, in theory, yield the most favorable performance for true off-target prediction, as the model would receive equal signal from both classes. However, this setting is not realistic for genome-wide CRISPR screening, where true off-target events are extremely rare compared to the number of potential genomic sites. Therefore, model performance must be evaluated under conditions of class imbalance that reflect real-world constraints. One way to study this is to evaluate the model across different class ratios while keeping the evaluation metrics focused on minority-class performance. As the ratio becomes more imbalanced, overall accuracy may increase, but metrics such as AUPRC, macro F1-score, and positive-class recall provide a more meaningful measure of whether the model can still reliably identify true off-target events. Under this framework, the objective is not to optimize performance at a single idealized ratio, but to assess how well the model maintains minority-class detection as the data distribution approaches realistic genome-scale conditions. The 5:1 ratio used in this work represents a compromise between these considerations. It substantially reduces the dominance of negative samples during training, allowing the model to learn meaningful patterns associated with true off-target events, while

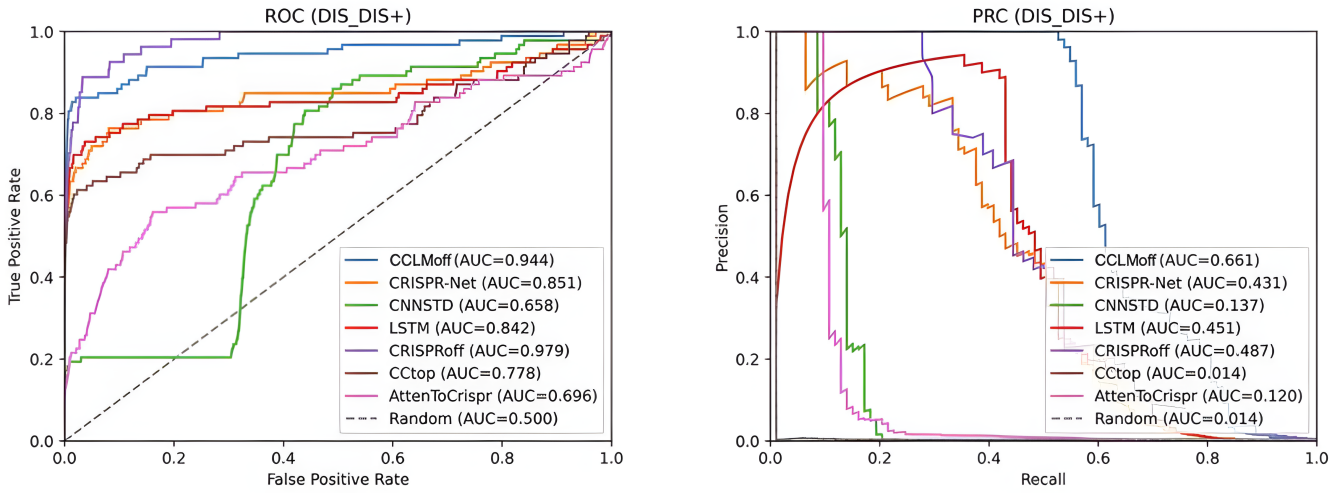

Figure. 5 Best models currently available and their results on the Leave One Dataset Out method using the GUIDE-seq dataset for evaluation [21].

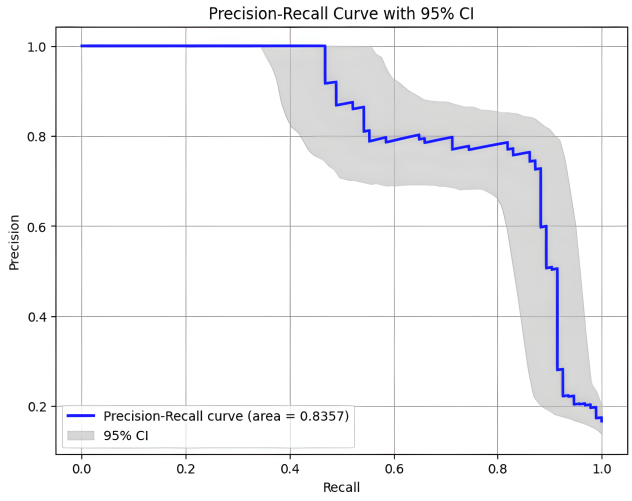

Figure. 6 Precision-Recall Curve for DIS-seq and DISplus-seq

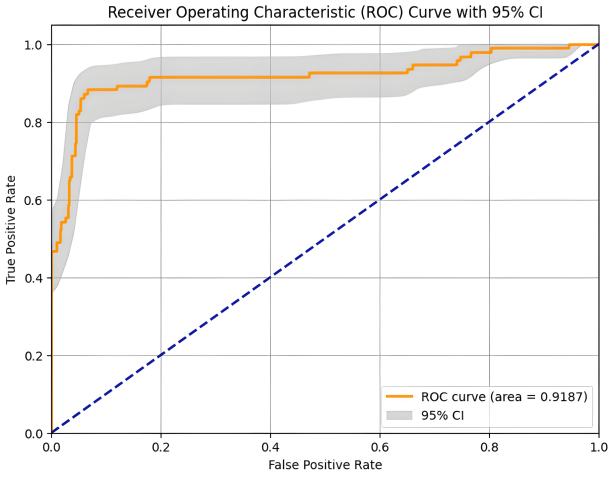

Figure. 7 Receiver Operating Characteristic Curve for DIS-seq and DISplus-seq. The dashed line indicates random chance.

still preserving an imbalanced setting that better approximates real screening data than a fully balanced dataset. Metrics such as AUPRC become particularly appropriate in this context, as they emphasize performance on the positive class and are not inflated by the large number of true negatives. Under this ratio-aware evaluation framework, the observed improvements in minority-class F1-score and AUPRC indicate that the model's performance gains are not simply artifacts of aggressive rebalancing, but instead reflect improved detection of biologically relevant off-target edits. This is especially important in practical applications, where false negatives pose a significantly greater risk than false positives. However, with a positive-class recall of 0.5592, approximately 44% of true off-target sites in the test set are still missed, and the model has not been evaluated at native class prevalence, across all assay holdouts, or with wet-lab confirmation. It should therefore be understood as an improved computational screening step rather than a clinically deployable classifier.

However, the model still has a few minor limitations. The first limitation is that random sampling is employed to gen-

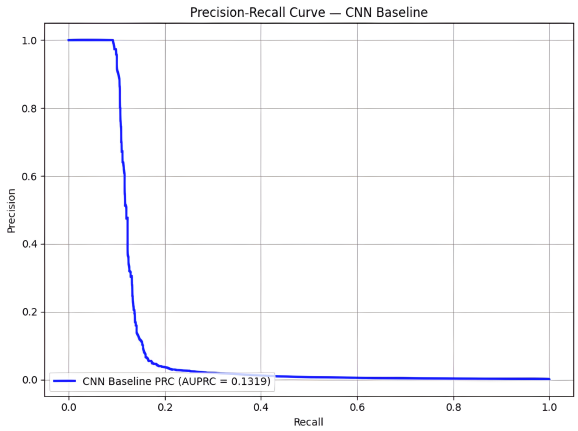

Figure. 8 Precision-Recall Curve for 1D CNN Baseline Model on GUIDE-seq

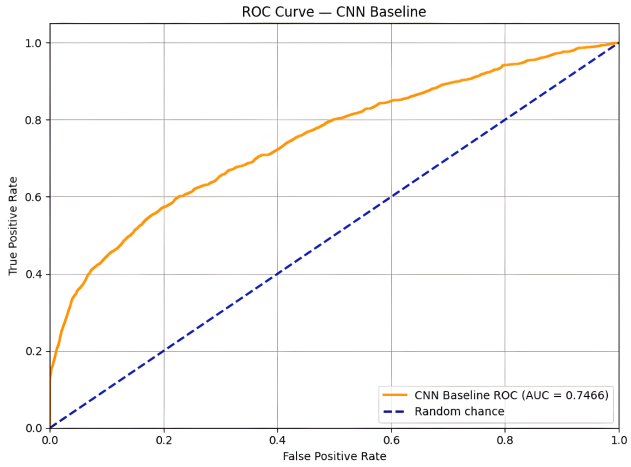

Figure. 9 Receiver Operating Characteristic Curve for 1D CNN Baseline Model on GUIDE-seq. The dashed line indicates random chance.

erate the training subsets. This random sampling is crucial to ensure that the model encounters multiple variations in the data. However, due to inherent randomness, it is also possible, albeit very unlikely, that insufficient variation exists to develop an accurate model. In this case, the training process would need to be rerun to resolve this issue, which could require additional compute resources and time. The second limitation is that some genome sequences in the dataset were simulated. These simulated sequences were generated using CasOFFinder [10], a high-level tool that simulates conditions closely approximating those in the real world. It is impractical to obtain sufficient numbers of real genome sequences for machine learning, as CRISPR technology is not yet widespread or sufficiently reliable. A further limitation concerns the variation in class imbalance ratios across the individual detection assays included in the training partition. The CCLMoff benchmark dataset contains assays with positive-to-negative ratios ranging from 1:26 to 1:4189, meaning that the leave-one-method-out evaluation is sensitive not only to sequence distribution shift between assays but also to their differing imbalance profiles. The present study does not systematically examine how this per-assay imbalance variation influences cross-assay generalization, and future work should evaluate model performance stratified by the imbalance ratio of the held-out assay to disentangle the effects of distributional shift from those of label scarcity.

Conclusion

This study evaluated the hypothesis that a computationally efficient distilled transformer encoder, combined with training set rebalancing, focal loss, and class weighting, would achieve higher AUPRC and minority-class F1-score than CCLMoff under an identical leave-one-dataset-out evaluation on the held-out GUIDE-seq test set. The hypothesis was supported: the proposed model achieved an AUPRC of 0.6157 and a positive-class F1-score of 0.5701, compared to 0.513 and 0.429 for CCLMoff under the same protocol. The CNN ablation further demonstrated that these gains are attributable to the transformer architecture rather than the imbalance-correction strategy alone, as a 1D CNN baseline trained under identical conditions collapsed to near-zero precision. Taken together, these findings establish that a lightweight general-purpose transformer, when paired with targeted imbalance-correction, can exceed the minority-class detection performance of a larger domain-pretrained model on this task.

Three specific directions for future investigation arise directly from the present findings. First, the sensitivity analysis across sampling ratios demonstrated that the 1:5 ratio outperforms more extreme imbalance settings; it remains untested whether a domain-pretrained genomic transformer such as DNABERT, fine-tuned under the same 1:5 rebalancing strategy, would further improve minority-class performance, or whether the gains observed here are specific to the character-level tokenization scheme used with DistilBERT. This is a directly testable hypothesis using the same evaluation protocol. Second, all results reported here are based on a resampled version of the CCLMoff benchmark; it is unknown whether the model's AUPRC and F1-score hold at native class prevalence across all available assay holdouts. Evaluating the model under a full leave-one-assay-out protocol at native prevalence would quantify the degree to which the rebalancing strategy inflates reported performance. More accurate off-

target edit detection reduces unnecessary experimental validation and lowers the risk of unintended genome edits, supporting safer use of CRISPR technology. Future work may incorporate additional biological context, such as the chromosome location and direction of the sequence, or evaluate the model on newly developed datasets to further improve generalization and robustness. In addition, the class imbalance ratio could be further optimized by running the model several times with different ratios. An optimized ratio would further improve the balance between real-world usability and model accuracy. This system can be improved further by creating an ensemble of models. Under this system, several such models are trained, and their predictions are combined to make a collaborative decision. This may further reduce the risk of error and optimize the model performance.

Acknowledgements

I would like to thank my parents and my mentor for their guidance and support throughout the process of this project and paper. Full code is accessible at: github.com/IshaanNirmal/CRISPR-Off-Target-Edit-Detection

References

1 The genetics and pathophysiology of single gene disorders in human. in Integrated Health and Sustainability Plants, Wildlife, and Genetic Resilience (eds R. Kausar, Z. U. Nisa, M. Jamil & I. Bashir) Unique Scientific Publishers, 2025. https://doi.org/10.47278/book.HH/2025.98.

2 M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier. A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. Science (New York, N.Y.). Vol. 337, pg. 816–821, 2012, https://doi.org/10.1126/science.1225829.

3 Recent therapeutic gene editing applications to genetic disorders — mdpi. https://www.mdpi.com/1467-3045/46/5/255?utm_source=chatgpt.com.

4 M. Asmamaw, B. Zawdie. Mechanism and applications of crispr/cas-9-mediated genome editing. Biologics: Targets & Therapy. Vol. 15, pg. 353–361, 2021, https://doi.org/10.2147/BTT.S326422.

5 Targeting specificity of the crispr/cas9 system — acs synthetic biology. https://pubs.acs.org/doi/10.1021/acssynbio.7b00270.

6 C. Guo, X. Ma, F. Gao, Y. Guo. Off-target effects in crispr/cas9 gene editing. Frontiers in Bioengineering and Biotechnology. Vol. 11, pg. 1143157, 2023, https://doi.org/10.3389/fbioe.2023.1143157.

7 Z. Sherkatghanad, M. Abdar, J. Charlier, V. Makarenkov. Using traditional machine learning and deep learning methods for on- and off-target prediction in crispr/cas9: a review. Briefings in Bioinformatics. Vol. 24, pg. bbad131, 2023, https://doi.org/10.1093/bib/bbad131.

8 H. Wang, Y. Wang, Z. Luo, X. Lin, M. Liu, F. Wu, H. Shao, W. Zhang. Advances in off-target detection for crispr-based genome editing. Human Gene Therapy. Vol. 34, pg. 112–128, 2023, https://doi.org/10.1089/hum.2022.198.

9 C. Li, W. Chu, R. A. Gill, S. Sang, Y. Shi, X. Hu, Y. Yang, Q. U. Zaman, B. Zhang. Computational tools and resources for crispr/cas genome editing. Genomics, Proteomics & Bioinformatics. Vol. 21, pg. 108–126, 2023, https://doi.org/10.1016/j.gpb.2022.02.006.

10 S. Bae, J. Park, J.-S. Kim. Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 rna-guided endonucleases. Bioinformatics (Oxford, England). Vol. 30, pg. 1473–1475, 2014, https://doi.org/10.1093/bioinformatics/btu048.

11 J. Listgarten, M. Weinstein, B. P. Kleinstiver, A. A. Sousa, J. K. Joung, J. Crawford, K. Gao, L. Hoang, M. Elibol, J. G. Doench, N. Fusi. Prediction of off-target activities for the end-to-end design of crispr guide rnas. Nature Biomedical Engineering. Vol. 2, pg. 38–47, 2018, https://doi.org/10.1038/s41551-017-0178-6.

12 J.-P. Concordet, M. Haeussler. CRISPOR: intuitive guide selection for crispr/cas9 genome editing experiments and screens. Nucleic Acids Research. Vol. 46, pg. W242–W245, 2018, https://doi.org/10.1093/nar/gky354.

13 J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten, D. E. Root. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. Nature Biotechnology. Vol. 34, pg. 184–191, 2016, https://doi.org/10.1038/nbt.3437.

14 S. Q. Tsai, Z. Zheng, N. T. Nguyen, M. Liebers, V. V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A. J. Iafrate, L. P. Le, M. J. Aryee, J. K. Joung. GUIDE-seq enables genome-wide profiling of off-target cleavage by crispr-cas nucleases. Nature Biotechnology. Vol. 33, pg. 187–197, 2015, https://doi.org/10.1038/nbt.3117.

15 T. J. Cradick, P. Qiu, C. M. Lee, E. J. Fine, G. Bao. COSMID: a web-based tool for identifying and validating crispr/cas off-target sites. Molecular Therapy. Nucleic Acids. Vol. 3, pg. e214, 2014, https://doi.org/10.1038/mtna.2014.64.

16 G. Chuai, H. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan, F. Gu, S. Qu, D. Huang, J. Wei, Q. Liu. DeepCRISPR: optimized crispr guide rna design by deep learning. Genome Biology. Vol. 19, pg. 80, 2018, https://doi.org/10.1186/s13059-018-1459-4.

17 H. K. Kim, Y. Kim, S. Lee, S. Min, J. Y. Bae, J. W. Choi, J. Park, D. Jung, S. Yoon, H. H. Kim. SpCas9 activity prediction by deepspcas9, a deep learning–based model with high generalization performance. Science Advances. Vol. 5, pg. eaax9249, 2019, https://doi.org/10.1126/sciadv.aax9249.

18 D. Wang, C. Zhang, B. Wang, B. Li, Q. Wang, D. Liu, H. Wang, Y. Zhou, L. Shi, F. Lan, Y. Wang. Optimized crispr guide rna design for two high-fidelity cas9 variants by deep learning. Nature Communications. Vol. 10, pg. 4284, 2019, https://doi.org/10.1038/s41467-019-12281-8.

19 J. Lin, Z. Zhang, S. Zhang, J. Chen, K.-C. Wong. CRISPR-net: a recurrent convolutional network quantifies crispr off-target activities with mismatches and indels. Advanced Science. Vol. 7, pg. 1903562, 2020, https://doi.org/10.1002/advs.201903562.

20 Q. Liu, D. He, L. Xie. Prediction of off-target specificity and cell-specific fitness of crispr-cas system using attention boosted deep learning and network-based gene feature. PLOS Computational Biology. Vol. 15, pg. e1007480, 2019, https://doi.org/10.1371/journal.pcbi.1007480.

21 W. Du, L. Zhao, K. Diao, Y. Zheng, Q. Yang, Z. Zhu, X. Zhu, D. Tang. A versatile crispr/cas9 system off-target prediction tool using language model. Communications Biology. Vol. 8, pg. 882, 2025, https://doi.org/10.1038/s42003-025-08275-6.

22 V. Sanh, L. Debut, J. Chaumond, T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019, https://doi.org/10.48550/arXiv.1910.01108.

23 T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár. Focal loss for dense object detection. in 2017 IEEE International Conference on Computer Vision (ICCV) pg. 2999–3007, 2017, https://doi.org/10.1109/ICCV.2017.324.

24 I. Loshchilov, F. Hutter. Decoupled weight decay regularization. in 2018.

25 C. Kim, S. Kim, J. Kim, D. Lee, S. Kim. Automated learning rate sched-

uler for large-batch training. Preprint at http://arxiv.org/abs/2107.05855 2021, https://doi.org/10.48550/arXiv.2107.05855.