# Quantifying uncertainty in off-target activity for CRISPR guide RNAs

Furkan Özden* and Peter Minary

Department of Computer Science, University of Oxford, Oxford OX1 3QD, United Kingdom

furkan.ozden@cs.ox.ac.uk, peter.minary@cs.ox.ac.uk

**Abstract.** CRISPR-based genome editing technologies have revolutionised the field of molecular biology, offering unprecedented opportunities for precise genetic manipulation. However, off-target effects remain a significant challenge, potentially leading to unintended consequences and limiting the applicability of CRISPR-based genome editing technologies in clinical settings. Current literature predominantly focuses on point predictions for off-target activity, which may not fully capture the range of possible outcomes and associated risks. Here, we present CRISPAI, a hybrid multitask neural network architecture approach for predicting uncertainty estimates for off-target cleavage activity, providing a more comprehensive risk assessment and facilitating improved decision-making in single guide RNA (sgRNA) design and experimental optimization. Our approach makes use of the count noise model Zero Inflated Negative Binomial (ZINB) to model the uncertainty in the off-target cleavage activity data. We show that uncertainty estimates of our approach are calibrated and its predictive performance is superior to the existing *in-silico* off-target cleavage activity prediction tools.

**Keywords:** Genome Editing · CRISPR · Uncertainty Quantification · Deep Learning.

## Introduction

CRISPR/Cas9 system (Clustered regularly interspaced short palindromic repeats/ CRISPR-associated protein 9) first discovered in the immune mechanisms of bacterial and archeal species (Bhaya et al., 2011) and quickly became among the most popular gene editing technologies recently with successful applications of editing eukaryotic genomes (Hsu et al., 2014). CRISPR/Cas9 system has also been applied to knockout-screening studies (Shalem et al., 2014), accelerating the understanding of variant and gene functions by uncovering causal

relations between mutations and phenotypes. Due to its high efficiency, simpler design and easier operation procedures in comparison to earlier genome editing methods like Zinc Finger Nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs), it is becoming a standard in genome engineering field and has the potential to lead new treatments for genetic diseases (Barrangou and Doudna, 2016).

The CRISPR/Cas9 system works by using a programmed single guide RNA (sgRNA) to direct the Cas9 enzyme to a specific target sequence in the targeted DNA site. Once the Cas9 enzyme is bound to the target DNA, it creates a double-strand break, which triggers natural repair mechanisms of the cell. This can result in the targeted sequence being edited by deletion, insertion, or replacement, depending on the desired outcome. However while CRISPR/Cas9 system operates on the targeted DNA region, cleavage may also occur at other genomic loci with a DNA that is not fully complementary to the sgRNA and has several base mismatched sites with the sgRNA. These cleavage effects, referred to as 'off-target' cleavage, are unintended and can be dangerous, resulting in unintended changes in the genome, leading to unwanted gene mutations and potentially harmful effects (Cho et al., 2014). The existence of the off-target cleavage phenomenon has been one of the key factors to limit the development and applicability aspects of the CRISPR-based genome editing systems. Studies found out that few-mismatch DNA sites are potentially recognizable by the sgRNA during the guiding process (Zhang et al., 2015). Also, the off-target effects have been shown to be dependent on many other factors such as nucleosome occupancy, chromatin accessibility and both binding and heteroduplex energy parameters of the sgRNA of choice (Mak et al., 2022). However, there are many potential sgRNA sequences that can direct the system for the intended cleavage effect to take place and hence one of the key design aspects of CRISPR-based systems is to evaluate and assess the error profile of the sgRNA of interest.

Up to date, many off-target cleavage activity prediction tools have been proposed to predict the potential off-target activity of a given target-sgRNA pair (i.e., targeted DNA site and corresponding single guide RNA sequences). These algorithms use the data generated by experimental CRISPR off-target detection assays, such as GUIDE-seq (Tsai et al., 2015), CHANGE-seq (Lazzarotto et al., 2020), DIGENOME-seq (Kim et al., 2015), CIRCLE-seq (Tsai et al., 2017), and predict a point score for off-target cleavage activity for a given target-sgRNA pair based on the training data. These tools can be divided into two main categories: (i) Conventional Machine Learning-based models and (ii) Deep Learning-based models. Conventional machine learning models have been

extensively used for on- and off-target prediction in CRISPR/Cas9. Various algorithms such as random forest, SVM (Doench et al., 2016), logistic regression (Chen and Tran, 2019), gradient boosting (Listgarten et al., 2018), and ensemble learning (Zhang et al., 2019) have been employed to predict off-target activity. While conventional machine learning models have shown promising results, recent studies using deep learning techniques have demonstrated even better performance (Liu et al., 2020). These models utilize novel sequence encoding strategies, feature engineering approaches by introducing physical features (Störtz et al., 2023), class rebalancing techniques, and attention mechanisms (Liu et al., 2019) to improve prediction performance. While several models have been developed, most of them are primarily dedicated to the classification task, aiming to predict the activity status of the sgRNA-target interface. Notably, CnnCrispr (Liu et al., 2020) stands out as a top-performing model which tackles the regression task of predicting the activity level of the sgRNA-target interface using only sequence-based features, which is the focus in this article as well. Sherkatghanad et al. (2023) presents a detailed overview of machine learning and deep learning-based studies for on/off-target activity prediction task for CRISPR/Cas9 systems.

It is worth noting that, the imbalance in CRISPR off-target prediction data poses a significant challenge, as the number of true off-target sites recognized by whole-genome detection techniques is much smaller than that of all possible nucleotide mismatch loci (Gao et al., 2020; Sherkatghanad et al., 2023). This imbalance can make training routine machine learning models difficult, resulting in high accuracies for the majority class but poor performance for the minority class, which is of greater interest in this context because it represents the actual off-target sites that can lead to unintended consequences in genome editing (Listgarten et al., 2018). These point predictions, while informative, may not fully capture the range of possible outcomes or the associated risks in the editing process. To the best of our knowledge, no such tool has focused on the uncertainty in off-target activity predictions. Incorporating uncertainty estimates into predictive models would facilitate the identification and prioritization of potential off-target sites, especially when they have similar point off-target activity predictions. By considering the uncertainty estimates, researchers and practitioners can differentiate between sites with similar point predictions and prioritize those with lower uncertainty, thereby reducing the chances of unforeseen off-target effects. This prioritization strategy, enabled by uncertainty estimates, would lead to improved validation and optimized guide RNA design, reducing potential risks associated with

CRISPR-based genome editing applications. Additionally, genome-wide off-target detection methods encounter various experiment-specific limitations that can affect their sensitivity. One of the key contributing factors responsible from the highly imbalanced nature of the off-target data produced by such detection techniques stems from the fact that these methods rely on Next Generation Sequencing (NGS)-based DNA sequencing assays which generally exhibit high sparsity, often encountered in microbiome, bulk-, and single-cell RNA experiments (Silverman et al., 2020; Eraslan et al., 2019). These limitations hinder the ability to differentiate between real biologically inactive off-target sites and technical errors, resulting in false negatives in the analysis for the guide RNA of interest (Naeem et al., 2020). Addressing the highly imbalanced nature, and the uncertainty, of the off-target cleavage data is crucial, particularly because mistaking an active off-target site for an inactive one can have significant consequences. Such errors can disrupt cellular function or confound experimental interpretation, whereas mistaking an inactive site for an active one may only necessitate designing another gRNA (Listgarten et al., 2018). In this work we present CRISPAI, to accurately predict the off-target cleavage activity in a probabilistic framework, allowing for quantification of the uncertainty in the predictions.

## Results

### Overview of CRISPAI

CRISPAI is a multitask, hybrid, deep learning architecture based on the count distribution Zero-Inflated Negative Binomial (ZINB), which accounts for highly imbalanced nature of the off-target cleavage data and models the off-target cleavage activity in a probabilistic framework (Methods). Our architecture uses a combination of bi-directional Long Short Term Memory (biLSTM) and Convolutional Neural Network (CNN) layers to extract sequence-based features of the target-sgRNA pair interface, which are encoded using a binary matrix encoding scheme (Lin et al., 2020). The architecture is designed to predict three parameters associated with ZINB, while training the network weights in a multitask fashion by simultaneously learning to predict the activity label of the samples (i.e., whether it is a detected off-target site by the detection assay). Hence, we were able to boost the accuracy of predicted parameters for the corresponding ZINB distributions by jointly optimizing the network weights to: i) optimize the likelihood of the detected cleavage activity values and ii) accurately classify the activity labels, associated with the target-sgRNA interface (Fig. 1).

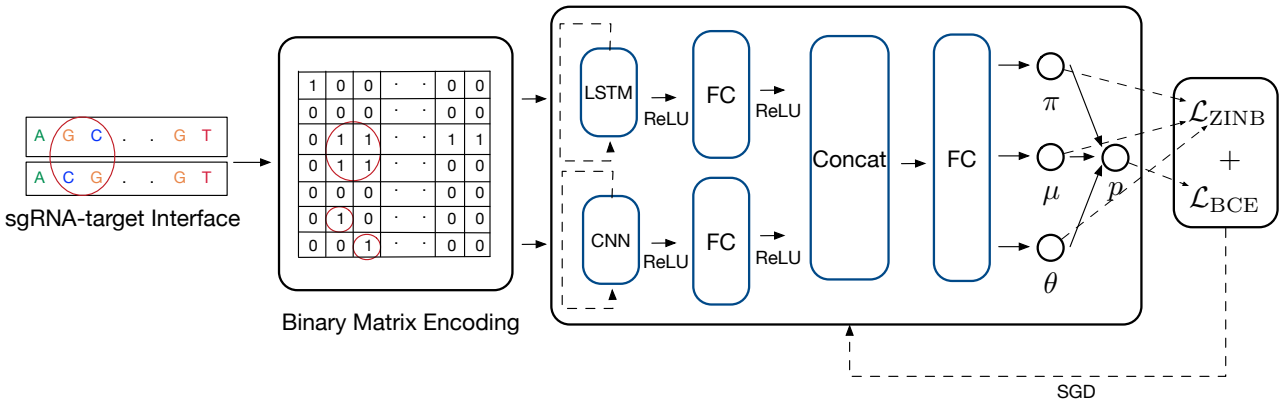


**Fig. 1.** Workflow Figure - A binary matrix encoding based on logical OR operation is employed to encode the target-sgRNA interface (Methods). Then, the encodings are passed to two parallel arms based on series of biLSTM and CNN layers. Three parameters $\pi$, $\mu$ and $\sigma$, are predicted with a Fully Connected (FC) layer using a concatenation of the extracted features by both arms. The predicted parameters are then inputted to another output neuron, which predicts the activity label of the sgRNA-target interface simultaneously, training the network weights in a multitask fashion.

### CRISPAI enables uncertainty quantification for off-target activity prediction

Our proposed architecture, CRISPAI, models the off-target cleavage activity in a probabilistic framework. Hence, it is able to sample from the posterior off-target activity distributions conditioned on the sequence-based features of the target-sgRNA interface. Fig. 2 depicts example distributions for eight randomly selected test samples from the CHANGE-seq (Lazzarotto et al., 2020) dataset, where point predictions and the ground truth CHANGE-seq detected frequencies are depicted with vertical lines.

We started evaluating CRISPAI by visualising the uncertainty estimates for individual target-sgRNA interface pairs on the separated test portion of CHANGE-seq dataset (Lazzarotto et al., 2020). CHANGE-seq is a highly scalable, next generation sequencing (NGS)-based *in-vitro* genome-wide Cas9 off-target detection assay which includes both epigenetic and genetic impact. Authors identified $202,043$ off-target sites for 110 sgRNAs on 13 therapeutically relevant loci in human primary T-cells. First, we obtained posterior parameters of the respective Zero-Inflated Negative Binomial distributions conditioned on the target-sgRNA interface features for each sample

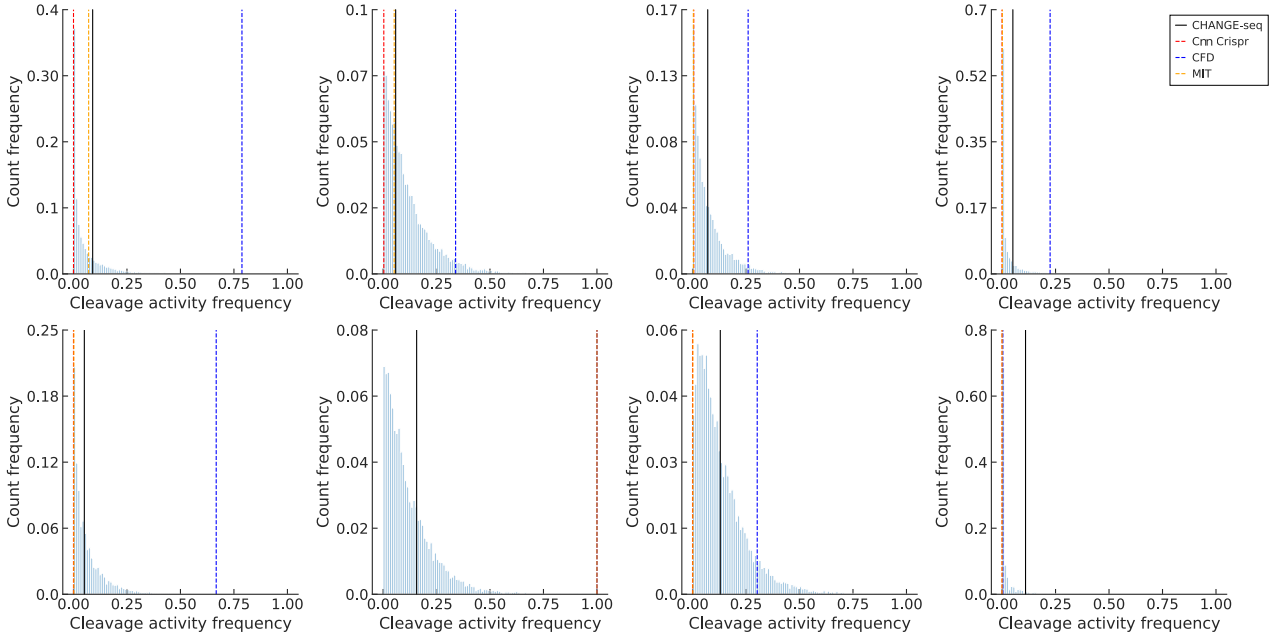


**Fig. 2.** Example distributions for eight randomly selected target-sgRNA interfaces. CRISPAI enables sampling from the posterior off-target activity distribution conditioned on the sequence-based features of the target-sgRNA pair of interest. Vertical lines represent ground truth CHANGE-seq, CnnCrispr, CFD score and MIT score detected off-target cleavage activities.

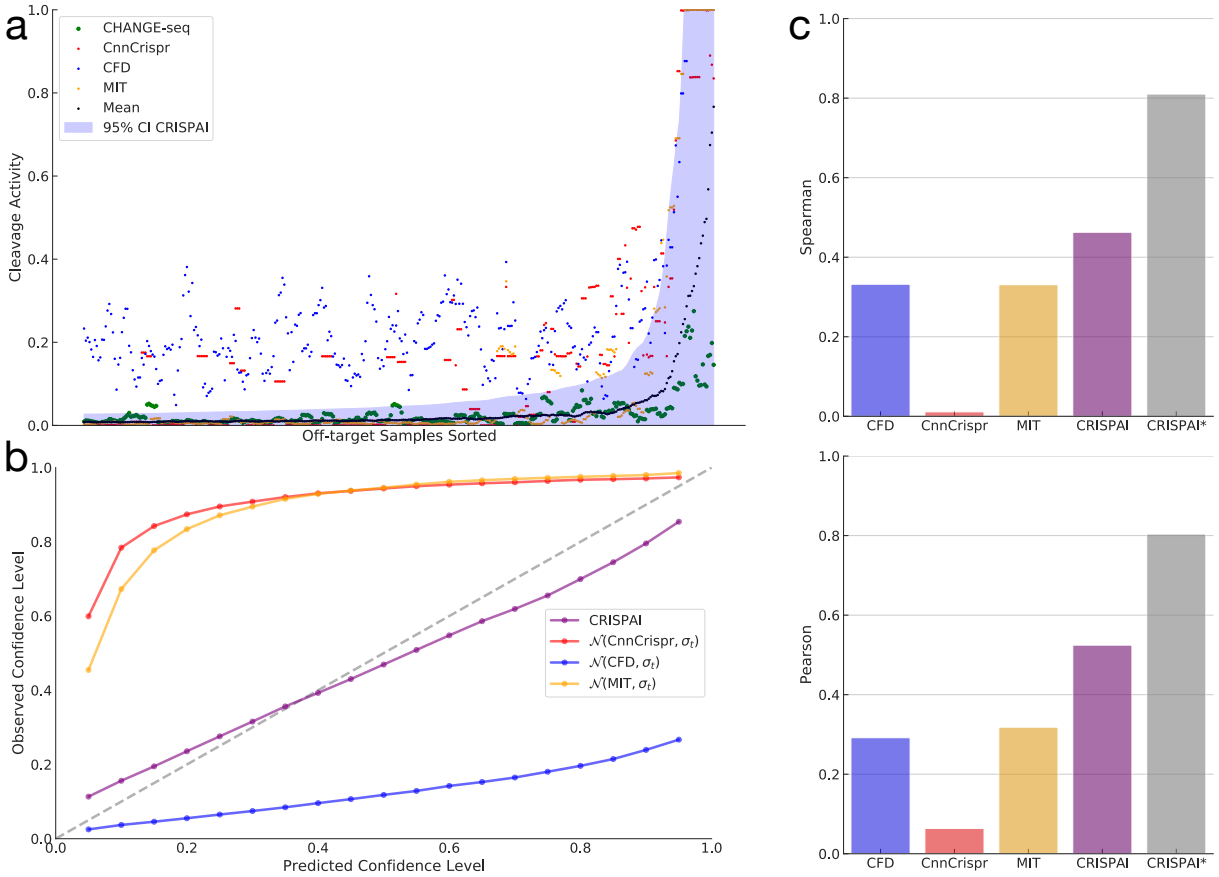


**Fig. 3.** Depiction of CRISPAI-predicted uncertainty estimates on test portion of the CHANGE-seq data. **a.** First, we obtained the CRISPAI-predicted off-target activity distributions for samples with positive CHANGE-seq counts. Then we calculated both the 95% confidence interval and the expected value for each distribution and sorted the off-target samples and associated predictions of CnnCrispr, CFD and MIT along with the ground truth CHANGE-seq detected activity (normalized between 0 and 1) with respect to the expected values in increasing order. **b.** For benchmarking purposes, we generated baseline uncertainty estimates by generating Normal distributions for each sample in the test set, $\mathcal{N}(\mu_i, \sigma^2)$, where $\mu_i$ and $\sigma^2$ are set to CnnCrispr, CFD and MIT prediction for the $i^{\text{th}}$ sample and variance of the test set respectively. **c.** We measured the Spearman and Pearson correlation coefficients of the ground truth CHANGE-seq detected activity and competing methods. We obtained the point off-target cleavage activity prediction with two different strategies using the CRISPAI-predicted activity distributions and the predicted activity label: i) taking the expected values of the obtained distributions and ii) using the predicted activity label value to assign zero activity for samples predicted as inactive while using the expected value for samples predicted as active (named CRISPAI*).

using CRISPAI. We sampled corresponding posterior distributions many times to obtain empirical probability mass functions (PMFs) for each sample. Ideally, in guide RNA design, a maximum cleavage activity on the targeted site with minimum activity on off-target sites is desired. Hence, Upper Confidence Bounds (UCB) of uncertainty estimates for the prediction of on-target activity task are not of concern and should be set to maximum activity value possible, measuring the Lower Confidence Bound (LCB) accordingly for the desired Confidence Interval (CI). Similarly, for LCB of uncertainty estimates for the prediction of off-target activity task are not of concern and should be set to lowest activity value possible - again, measuring the UCB for off-targets accordingly for the desired CI. Therefore, we set LCBs of all off-target samples at minimum activity (i.e., 0) and we measured the UCB at 95% CI. Fig. 3**a.** shows the confidence intervals of CRISPAI-predicted posterior distributions for off-target sites. We observed that predicted intervals accurately captured the ground truth cleavage activity frequency values for almost all of the samples. For on-target sites and highly active off-target sites (e.g., activity frequency $> 0.3$) the posterior distributions yielded very high 95% CI UCBs reaching up-to maximum activity frequency, which is expected since the detected frequency value highly depends on the number of other detected off-target sites for the same guide RNA instead of the individual features of the interface itself for highly active off-target sites and on-target sites (Lazzarotto et al., 2020). Additionally, we ran CnnCrispr (Liu et al., 2020), CFD score (Doench et al., 2016), and MIT score (Hsu et al., 2013) on the same data and obtained point predictions for comparison with the expected values of the CRISPAI-predicted posterior distributions for the same samples. We observed that the expected values for each off-target sites closely resemble the ground truth activity values, whereas CnnCrispr and CFD scores are generally far off while MIT score is somewhat better at distinguishing between minimal activity and highly active sites.

Next, to evaluate the quality of the uncertainty estimates of CRISPAI using the same test portion of the CHANGE-seq dataset, we computed the diagnostic calibration measure for uncertainty estimates proposed by Kuleshov et al. (2018). Proposed diagnostic measure suggests that a well-calibrated uncertainty forecast should contain $N-$percent of samples in $N-$percent confidence interval. By plotting the observed confidence level against the expected confidence level, we produced the proposed diagnostic plot, in which well-calibrated uncertainty estimates are expected to produce a straight line. For the sake of comparison with baselines, we created Normal distributions for each off-target site, centered at the point predictions of CnnCrispr, CFD and

MIT with standard deviations equal to the standard deviation of the test set. Figure 3**b.** depicts diagnostic lines for each aforementioned method. We observed that diagnostic plot of CRISPAI-predicted CI levels is similar to the ideal calibration line, while constructed base-line estimates are far off.

Figure 3**c**. shows the measured Spearman and Pearson correlations of CnnCrispr, CFD, MIT and CRISPAI on the same test portion. In addition to expected values, to obtain point predictions from CRISPAI-predicted posterior off-target activity distributions, we made use of the multitask design of our model to generate an alternative point prediction. We used the predicted active/inactive (Methods) classification scores to assign zero cleavage activity for samples predicted as inactive while we kept the predicted expected value for the samples predicted as active. We refer to expected value-only predictions as CRISPAI and CRISPAI* for the latter point predictions in Fig. 3**c**. We observed that both predictions significantly outperformed other tools with 63% and 150% improvements over, the best model for this test set, MIT score for CRISPAI and CRISPAI* respectively. The inferior performance of the CnnCrispr model compared to heuristic-based models (i.e., CFD and MIT) on a specific task highlights the vulnerability of task-specific optimized deep learning models to dataset shifts and generalizability issues. This leads to poorer performance in real-world applications beyond the dataset used for model development. To conduct a more detailed comparison with CnnCrispr, we evaluated our model on the same training and testing dataset used for the development of the CnnCrispr model. The results of this evaluation demonstrated that our tool outperforms CnnCrispr on this dataset as well (Fig. 4). We concluded that CRISPAI-predicted posterior off-target activity distributions provide accurate uncertainty estimates while not sacrificing and even greatly improving the predictive performance as we further validate with other test sets in the next section.

### CRISPAI improves *in-silico* CRISPR/Cas9 off-target cleavage activity prediction performance

The dataset curated by the authors of DeepCrispr consists of two cell-lines, namely HEK293t and K562 (Methods). The authors of CnnCrispr applied a $80\% - 20\%$ train-test split on this dataset and evaluated CFD, MIT, DeepCrispr and CNN_std on the test portion of the dataset with respect to various performance metrics for both classification and regression tasks for each of the cell-lines and also for the total dataset containing the test data from both of the cell-lines. To further evaluate the predictive performance of CRISPAI, we applied the

same train-test split ratio on the same dataset and we trained a CRISPAI model on the training portion. We evaluated our models performance on the test portion of the data using: (i) point predictions obtained through taking expectations of the predicted posterior off-target cleavage activity distributions; and (ii) using both the inactive/active class label and the expected values together as described in Fig. 3.

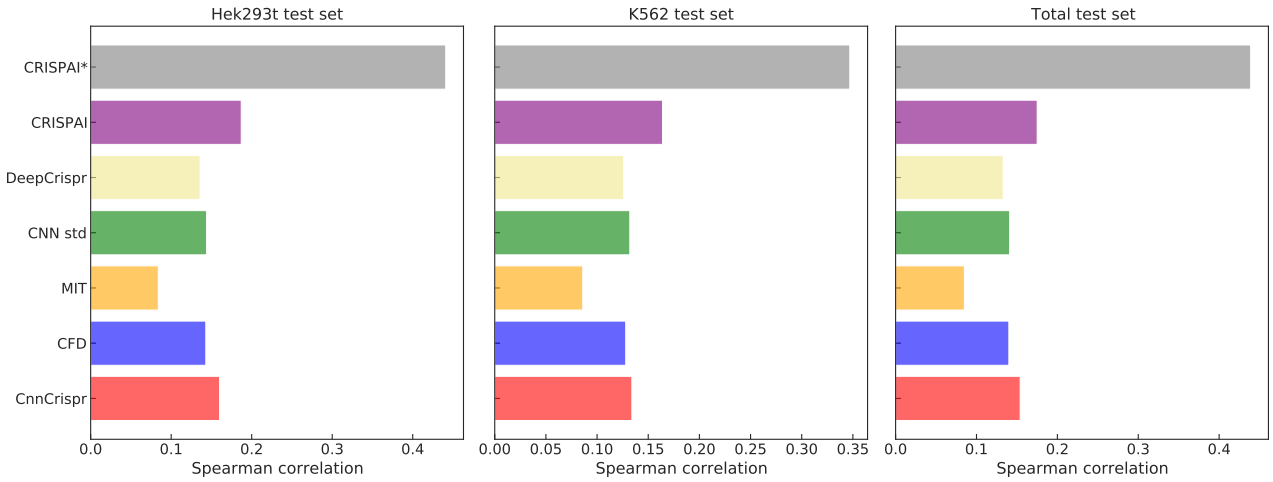


**Fig. 4.** Predictive performance comparison of competing models on the test dataset curated by Chuai et al. (2018). Bar-plots represent Spearman correlation coefficient between the ground truth cleavage activity and the predictions of DeepCrispr, Cnn_Std, MIT, CFD, CnnCrispr and CRISPAI. We measured the performance on 20% test portion of three datasets in total, similarly with CnnCrispr, where first two datasets are on Hek293t and K562 cell-lines and the total test set is the aggregation of the two cell-lines. Both CRISPAI (purple) and CRISPAI* (gray) point predictions significantly outperformed CnnCrispr (red), the best performing model, in all of the three testing scenarios.

We observed that both point predictions obtained using the distributions by our model, CRISPAI and CRISPAI*, significantly out-performed all of the competing tools with respect to Spearman correlation and performed very similar with respect to auROC where all competing tools, except MIT score, performed almost perfectly (Supplementary Table 1). Specifically, we observed 13% and 1.85 fold improvements in Spearman correlation value for the total test set over CnnCrispr, the best model in this category, from 0.154 to 0.175 and 0.439 for CRISPAI and CRISPAI* respectively. We observed similar trends for HEK293t and K562 test sets with 16% and 22% improvements for CRISPAI and 1.76 and 1.37 fold improvements for CRISPAI* over CnnCrispr, again the best model in this category, respectively for the two cell-lines. Other than CRISPAI and CRISPAI* Deep

learning-based models performed similarly with 0.154, 0.141, 0.133 Spearman correlation values for CnnCrispr, CNN_std and DeepCrispr respectively for total test set. Interestingly, CFD score performed comparable to Deep-learning based models unlike MIT score for this test set with 0.140 and 0.085 correlations respectively.

**CRISPAI uncovers effects of mismatches and off-target activity on the uncertainty of the predictions.**

To investigate whether CRISPAI-predicted off-target activity distributions capture biological principles behind CRISPR/Cas9 system, we used the test portion of CHANGE-seq dataset and stratified the samples with: (i) the base-pair mismatch count between the target site and sgRNA; and (ii) detected CHANGE-seq frequency and plotted the stratified folds with respect to the predicted variances of the associated off-target activity distributions.

We observed an almost linear decrease in the maximum variance of predicted distributions as the number of mismatches decreased from 6 to 0 (Fig. 5**a**). This result is expected since, as the number of allowed mismatches between target and sgRNA sequences increases, allowed degree of freedom for other sequence-based factors, that are shown to be correlated with higher off-target cleavage activity, increases (e.g., GC content, mismatch location) yielding a higher variance in predicted off-target activity distributions.

We observed a similar trend for the CHANGE-seq detected frequency (Fig. 5**b**), where the detected frequency values for samples with low variance predicted distributions are higher than those with distributions having higher variance. The inverse relationship between off-target cleavage activity and the variance of predicted distributions in our observed graph presents a valuable insight that supports the efficacy of our model. The fact that the detected off-target cleavage activity is inversely related to the variance implies that as off-target cleavage activity decreases, our model's confidence in predicting sgRNA-target interfaces increases. This phenomenon is highly advantageous as it indicates that our model is more certain and consistent in identifying sgRNA-target interfaces with higher off-target cleavage activity. These interfaces exhibit more densely distributed predicted probability distributions, suggesting a greater level of confidence in their predictions. This finding underscores

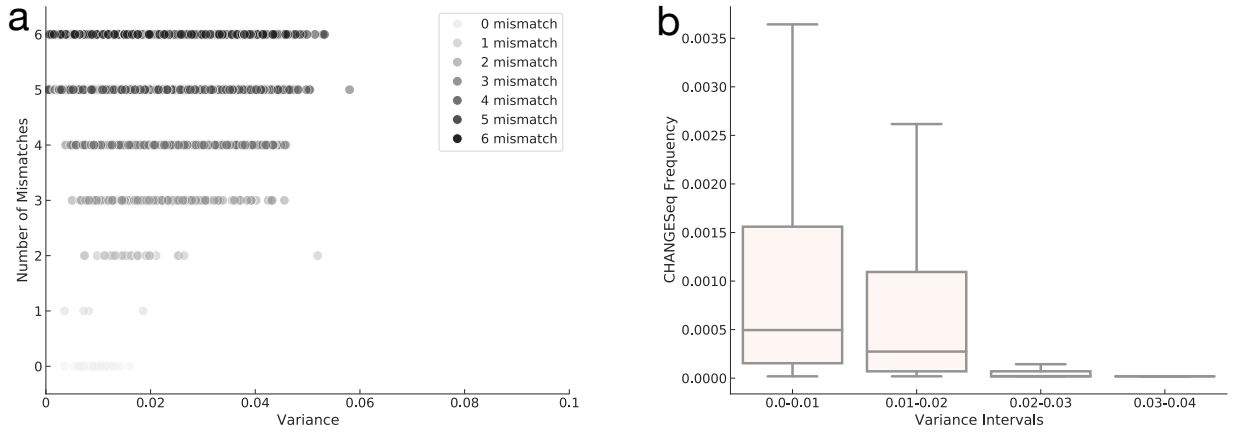


**Fig. 5.** Variance of CRISPAI-predicted off-target activity distributions on the test portion of the CHANGE-seq dataset. The figure depicts the relationship between base-pair mismatch count, detected CHANGE-seq frequency, and the predicted variances of associated off-target activity distributions. **a.** The graph shows the decrease in maximum variance as the number of mismatches decreases from 6 to 0. **b.** Similarly, the inverse relationship between detected frequency and the variance of predicted distributions is illustrated, indicating higher confidence in CRISPAI's predictions for sgRNA-target interfaces with lower off-target cleavage activity.

the reliability and robustness of our model, enhancing its potential for accurate and efficient identification of target sites with higher off-target cleavage activity.

## Discussion

The development of *in-silico* predictive models for CRISPR/Cas9 off-target activity prediction has achieved significant milestones in various approaches, including heuristic models, traditional machine learning models and deep learning models. Heuristic models were among the early attempts to predict off-target activity (Hsu et al., 2013; Doench et al., 2016; Stemmer et al., 2015; Montague et al., 2014), relying on predefined rules and sequence patterns. While these models provided initial insights, they often lacked generalizability and accuracy. The advent of learning models, such as traditional machine learning algorithms like SVM, random forest, and logistic regression, brought improvements by leveraging data-driven approaches (Wang et al., 2014; Doench et al., 2014; Xu et al., 2015; Listgarten et al., 2018). These models incorporated features derived from sequence characteristics and demonstrated enhanced prediction capabilities. However, with the emergence of deep learning models in the field, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the effectiveness and prediction capabilities of the modelling efforts significantly improved (Chuai et al., 2018; Lin et al., 2020; Liu et al., 2020; Zhang et al., 2020). Deep learning models could effectively handle large volumes of complex data, capturing intricate patterns and achieving superior performance in predicting off-target activity. The development of deep learning-based prediction models also enabled utilization of physical features modelling the off-target activity problem with more depth (Störtz et al., 2023). Accurately quantifying uncertainty in off-target activity predictions is a crucial next milestone this study aims to address. While predictive models have shown promising results in identifying potential off-target cleavage activity, they provide point predictions without considering the associated uncertainty. Incorporating uncertainty estimates into these models would provide a more comprehensive understanding of the reliability and confidence of the predictions. It would enable researchers and practitioners to differentiate between sites with similar point predictions but different levels of uncertainty, allowing for better risk assessment and prioritization of potential off-target sites. Additionally, quantifying uncertainty would enhance the transparency and communication of the prediction results, providing stakeholders with a clearer understanding of the associated risks.

Our approach quantifies the confidence in predicted off-target activities, guiding researchers in refining experimental conditions, selecting sgRNA sequences with lower uncertainty, and prioritizing experimental validation for potential off-target sites. Additionally, providing the research community and practitioners of CRISPR technology with uncertainty estimates in off-target prediction activity would be helpful in driving precision and reducing potential risks associated with gene editing applications. These estimates would enable scientists to better evaluate the likelihood of off-target effects, allowing for more informed decision-making in the design and execution of gene editing experiments. Furthermore, uncertainty estimates would facilitate the identification and prioritization of potential off-target sites, even when they have similar point off-target activity predictions. Hence, by considering the uncertainty estimates, researchers can differentiate between sites with similar point predictions and prioritize those with lower uncertainty, thereby reducing the chances of unforeseen off-target effects. This prioritization strategy would lead to improved experimental validation techniques and optimized guide RNA design, ultimately enhancing the overall reliability and safety of CRISPR-based applications.

For sequencing data analysis, it is essential to acknowledge the heterogeneity of zeros, as they can originate from diverse processes, introducing noise and uncertainty into the data. Thorough modeling is required to address this phenomenon and ensure accurate interpretation of results. Two primary categories of zeros are encountered: technical zeros and biological zeros (Silverman et al., 2020). Technical zeros arise from limitations in sample preparation or sequencing, leading to partial or complete reduction in countable sequences. Examples include biases in amplification, sequencing depth limitations, or batch effects. In contrast, biological zeros occur when a specific sequence is genuinely absent from the biological system under investigation, such as unique bacterial strains in different individuals or gene deletions in knockout experiments. Extensive research efforts have been devoted to overcoming this type of challenge in various application domains. Researchers have developed sophisticated techniques, including the use of technical components like zero-inflated distributions, to effectively model noisy zero counts. These approaches have been applied in domains such as microbiome (Weiss et al., 2017) studies, single-cell RNA sequencing (Eraslan et al., 2019; L Lun et al., 2016), and bulk RNA sequencing (Zhu et al., 2018), enabling improved accuracy and insight in differential gene expression analysis (Soneson and Robinson, 2018) and facilitating a deeper understanding of biological processes. In off-target activity detection data, we use technical zeros to refer to the off-target sites that can not be captured due to

the limited sensitivity of the detection assays. To address the technical zero problem in the raw count version of off-target activity data, we employed a modeling approach that utilized a Zero-Inflated Negative Binomial (ZINB) distribution. This allowed us to accurately model the characteristics of the data, account for excessive zeros, and incorporate uncertainty modeling.

Although our approach provides a more comprehensive risk assessment than traditional point predictions, it is important to note that our modelling effort tackles the uncertainty problem by considering abundant number of potential off-target sites, which is not suitable for modelling uncertainty in the on-target cleavage activity. The current model shares the limitation of all sequence-based models, whose predictions are solely based on the sgRNA-target DNA sequence pair because they cannot differentiate between off-target activities of identical sgRNA-target interfaces at different genomic loci. However, this limitation can be circumvented by including additional features (e.g., characterising to epigenetics and primary chromatin structure) capturing the biological environment at the genomic loci as this was done in our earlier work (Störtz et al., 2023). As future work, the next challenges will be developing suitable models for modelling uncertainty in the on-target cleavage activity and including features that capture the biological environment at the off-target sites.

# Methods

## Datasets

We conducted three different training for all the testing scenarios described in the Results section, to evaluate CRISPAI's performance, namely: Training on (i) $202,043$ CHANGE-seq (Lazzarotto et al., 2020) detected off-target sites on T-Cells for 110 sgRNAs; (ii) $1,703$ GUIDE-seq detected off-target sites for the same sgRNAs (Supplementary Figure 1); and (iii) 656 GUIDE-seq, Digenome-seq, BLESS, HTGTS, and IDLV detected off-target sites curated and presented by the authors of DeepCRISPR (Chuai et al., 2018) on HEK293T and K562 cell-lines for 18 and 12 sgRNAs respectively. The above mentioned numbers refer to detected, which we refer to as *active*, off-target sites, whereas there are much more potential off-target sites for each sgRNA which can not be detected due to limited sensitivity of the detection assays. However, many studies (Manghwar et al., 2020; Martin et al., 2016; Cho et al., 2014; Pattanayak et al., 2013) suggest that several mismatch (e.g., up-to 6 base-pairs) target sequences with respect to the sgRNA sequence are potential, which we refer to as *putative*, off-target sites. Also, we refer to putative off-target sites that are not active as *inactive* off-target sites. Hence, there arises two tasks for a given off-target sgRNA pair: Classification into active and inactive classes; and the regression task of predicting the off-target cleavage activity.

Many genome alignment-based methods (Montague et al., 2014; Bae et al., 2014; Concordet and Haeussler, 2018) have been proposed for discovery of the putative off-target sites for a given sgRNA. We employed one of the most popular, light-weight tool CasOFFinder (Bae et al., 2014) for this task owing to its ease of use and search speed. Specifically, we searched for putative off-target sites for all 110 sgRNAs presented in CHANGE-seq dataset with up to 6 base-pair mismatches with the sgRNA sequence. We obtained a total of $1,783,801$ putative sites, yielding $1,581,757$ inactive off-target sites for CHANGE-seq detected active off-target sites. As for the DeepCRISPR dataset, authors used the software *bowtie2*, originally developed for alignment of short DNA sequences to a reference genome (Langmead and Salzberg, 2012), and reported $\sim 160,000$ putative off-target sites for all 30 sgRNAs. It is also worth noting that, obtained datasets are highly imbalanced with imbalance ratios (i.e., number of negative samples per positive sample) 7.82 and 232.59 for CHANGE-seq and DeepCRISPR datasets respectively.

## CRISPAI Model

**Problem formulation.** Letting $x \in \{0,1\}^{m\times\ell}$, where $m$ and $\ell$ are feature dimension and sequence length respectively, be the nucleotide sequence-based features of sgRNA-target pair interface, $z \in \{0,1\}$ denote whether the off-target is inactive or active and $y \in \mathbb{R}$ denote the cleavage read depth, we model off-target activity data generation process with a Zero-Inflated Negative Binomial (ZINB) distribution. The ZINB distribution is an appropriate choice for modeling count data that is both highly sparse and overdispersed. A ZINB mixture model can be constructed using two components: a point mass at zero, which represents the excessive number of undetected inactive samples in the data, and a negative binomial component that models the count distribution. In the context of off-target assay data for CRISPR-based editing technologies, the point mass at zero is expected to capture the abundance of undetected inactive samples, while the negative binomial component is used to represent the sequencing reads for active samples. Hence, we model $y$ with the following set of parametric equations as:

$$\begin{aligned} \mathrm{NB}(k;\mu,\theta) &= \frac{\Gamma(k+\theta)}{\Gamma(\theta)}\left(\frac{\mu}{\mu+\theta}\right)^k\left(\frac{\theta}{\mu+\theta}\right)^\theta, \\ \mathbb{P}(y|x) &\sim \mathrm{ZINB}(k;\pi,\mu,\theta) = \pi\delta_0(k) + (1-\pi)\mathrm{NB}(k;\mu,\theta) \end{aligned} \tag{1}$$

$$\begin{aligned} f_{\mathrm{w}} &: \{0,1\}^{m\times\ell} \to \mathbb{R}^4, x \to (\pi,\mu,\theta,p), \\ p &= g_{\mathrm{w}'}(\pi,\mu,\theta), \quad \mathrm{w}' \subset \mathrm{w} \end{aligned} \tag{2}$$

where $\mathrm{ZINB}(k;\pi,\mu,\theta)$ is the ZINB distribution with parameters $\pi$, $\mu$ and $\theta$ for the mixture coefficient representing the point-mass ($\delta_0$) at 0, mean and dispersion of the negative binomial component (NB), respectively. We model the parameters for the conditional distribution with a multi-output parametric function $f_{\mathrm{w}} : \{0,1\}^{m\times\ell} \to \mathbb{R}^4$ with parameter set w. Note that the output space of the function $f_{\mathrm{w}}$ is $\mathbb{R}^4$, where the last dimension is reserved to predict a score for $z$ (classification task), with the parametric functional mapping $g_{\mathrm{w}'}$ using a proper subset of parameters $\mathrm{w}' \subset \mathrm{w}$ and utilizing advantages of multi-task training which are discussed in the CRISPAI architecture section. Thus, using a data-set consisting of $N$ samples in the form of 3-tuples, $\mathcal{D} = \{(x_i, y_i, z_i)\}_i^N$, we want to be able to calculate the posterior distribution of the cleavage score $y$ given the features $x$ and the data $\mathcal{D}$ — that is $\mathbb{P}(y|x,\mathcal{D})$.

**Target-sgRNA interface encoding.** We used 4-bit one-hot-encoding vectors representing the letters in the alphabet $\{A, G, C, T\}$. Using the one-hot-encoded representations for each base in sgRNA and target sequences yields $4 \times 23$ binary matrices for each sequence. Then we employed the target-sgRNA pair encoding scheme proposed in (Lin et al., 2020). We represent any sgRNA-target sequence pair with a $6 \times 23$ matrix as follows: First, both sgRNA and the target sequence is one-hot encoded. Then, obtained binary matrices are merged via an element-wise OR operation. Hence, resulting $4 \times 23$ binary matrix shows the mismatches between sgRNA-target sequence pair. However, OR operation does not preserve the direction of the mismatch. To help ameliorate this information loss, a two-bit direction channel is concatenated to the resulting binary matrix. For example, at a base-pair loci, '$0011 - 10$' represents the mismatch '$G \rightarrow C$'; '$0011 - 01$' represents the mismatch '$C \rightarrow G$' and one-hot vector '$0100 - 00$' represents the matched loci '$T \rightarrow T$'.

**CRISPAI architecture and training.** CRISPAI is an end-to-end multi-output CNN and bi-LSTM fusion neural network specifically designed to classify the active/inactive status and to quantify the uncertainty in off-target cleavage activity of target-sgRNA pairs for CRISPR/Cas9 system. We show that CRISPAI architecture is able to increase performance on off-target cleavage activity prediction task on different test sets, while providing useful uncertainty quantification (Results).

CRISPAI architecture is designed to estimate the parameters of the ZINB distribution conditioned on the target-sgRNA interface features. First, we input binary matrix encoding of the interface to a series of Convolutional Neural Network (CNN) and bi-directional Long-Short Term Memory (bi-LSTM) layers simultaneously for both spatial and temporal feature extraction. Specifically, we use 2 consecutive CNN layers with 128 and 32 kernels with 1 and 3 filter sizes respectively and a bi-directional LSTM layer with 128 hidden neurons in each direction. The output activations of the CNN layers are batch-normalized while the output of the last CNN layer is pooled with a MaxPool layer of kernel size 2 and with a stride of 2. The obtained encodings are further processed with two distinct 128-neuron Fully Connected (FC) layers individually. We then concatenate the final encodings outputted by the FC layers into a single 256-dimensional vector. The final FC layer with 64 neurons processes the concatenated vector and is inputted to 3 output neurons, one for each parameter of the ZINB distribution. We used ReLU activations for FC and CNN layers. The activation function choices for the individual output neurons predicting 3 associated parameters for the ZINB distribution, $\pi$, $\mu$ and $\theta$, are discussed below.

Studies have specifically reported that introducing inactive samples in the training of activity prediction models increases model performance for off-target cleavage activity prediction models due to additional information posed by the inactive samples (Yaish et al., 2022; Zhang et al., 2020). Hence, to reinforce the training of the neurons predicting these three parameters, we leveraged the additional information provided by the activity label of each interface. To achieve this, we used an additional output neuron which inputs the predicted parameters conditioned on the interface features and outputs a score for activity status of the respective target-sgRNA interface (Figure 1). The architecture can be formulated through following equations:

$$
\begin{aligned}
E_{cnn} &= \mathrm{ReLU}(\mathrm{CNN}(X)), \\
E_{bilstm} &= \mathrm{ReLU}(\mathrm{BiLSTM}(X)), \\
E_{total} &= \mathrm{Concat}(\mathrm{ReLU}(\mathrm{FC}_1(E_{cnn})), \mathrm{ReLU}(\mathrm{FC}_2(E_{bilstm}))), \\
\pi &= \mathrm{Logit}(\mathrm{FC}_\pi(\mathrm{ReLU}((\mathrm{FC}_3(E_{total}))))), \\
\mu &= \mathrm{Exp}(\mathrm{FC}_\mu(\mathrm{ReLU}((\mathrm{FC}_3(E_{total}))))), \\
\theta &= \mathrm{Exp}(\mathrm{FC}_\theta(\mathrm{ReLU}((\mathrm{FC}_3(E_{total}))))), \\
p &= \mathrm{Sigmoid}(\pi, \mu, \theta)
\end{aligned}
\tag{3}
$$

where $E_{cnn}, E_{bilstm}, E_{total}$ represent CNN encoding, biLSTM encoding and concatenated total encodings respectively. To ensure positivity on parameters $\pi$ and $\mu$, we use exponential activations, and for drop-out parameter $\pi$, we use Logit activation instead of Sigmoid for ease of integration with Negative Log Likelihood (NLL) loss for ZINB distribution.

All network parameters in (1), are learned with a multi-task training framework using Stochastic Gradient Descent (SGD) algorithm. To train the network weights, we use Negative Log Likelihood Loss of ZINB ($\mathcal{L}_{\mathrm{ZINB}} = \mathrm{NLL}_{\mathrm{ZINB}}$) and Binary Cross Entropy Loss ($\mathcal{L}_{\mathrm{BCE}}$) for three parameters of the ZINB distribution and

output neuron $p$ respectively.

$$\mathcal{L}_{\text{ZINB}}(\pi, \mu, \theta, y) = -\sum_{i=1}^{N} \log\left(\pi\delta_0(y_i) + (1-\pi)\text{NB}(y_i; \mu, \theta)\right),$$

$$\mathcal{L}_{\text{BCE}}(z, p) = -\frac{1}{N}\sum_{i=1}^{N}\left[z_i \log(p_i) + (1-z_i)\log(1-p_i)\right], \tag{4}$$

$$\mathcal{L}_{\text{total}}(\pi, \mu, \theta, z, p, y) = \mathcal{L}_{\text{BCE}}(z, p) \;\; + \;\; \mathcal{L}_{\text{ZINB}}(\pi, \mu, \theta, y)$$

We sum the individual losses into a cumulative loss function for each batch, and we apply SGD on the total loss function, $\mathcal{L}_{\text{total}}$. All parameters in (4) are defined in Problem Formulation section. Although a carefully tuned weighting on individual losses could improve training stability and performance (Kendall et al., 2018), we skipped this further optimization for the scope of this work.

We trained the CRISPAI architecture with the same training settings on all datasets. We splitted CHANGE-seq and GUIDE-seq dataset into training, validation and testing sets with 70%, 20% and 10% respectively. For DeepCRISPR data we did not use a validation set and used the same hyper-parameter configurations we discovered using the CHANGE-seq validation set. Hence, we applied a 80%-20% train-test split on the latter.

We used Adam optimizer (Kingma and Ba, 2014) for optimization with a learning rate of 0.00001. We stopped the training with early stopping, watching either the validation or training loss for 50-epochs, with a maximum epoch number of 500. We implemented CRISPAI on Python 3.9 using PyTorch (Paszke et al., 2019). Finally, we used SCVI-tools (Gayoso et al., 2022) to implement the loss functions. The models are trained on NVIDIA TITAN RTX GPUs.

## Code and Data Availability

All necessary scripts and data to replicate the results presented in the figures and tables, the tool and the trained models are available at https://github.com/furkanozdenn/crispai.

## Competing Interests

Authors declare no competing interest.

## Author Contributions

FO and PM designed the study. FO designed and implemented the model and performed the experiments. FO and PM wrote the manuscript.

## Rights Retention Statement

Authors acknowledge the funding from Google DeepMind. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

# Bibliography


Bae, S., Park, J., and Kim, J.-S., 2014. Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 rna-guided endonucleases. *Bioinformatics*, **30**(10):1473–1475.

Barrangou, R. and Doudna, J. A., 2016. Applications of crispr technologies in research and beyond. *Nature biotechnology*, **34**(9):933–941.

Bhaya, D., Davison, M., Barrangou, R., et al., 2011. Crispr-cas systems in bacteria and archaea: versatile small rnas for adaptive defense and regulation. *Annual review of genetics*, **45**(1):273–297.

Chen, S.-a. A. and Tran, E., 2019. Optimizing precision genome editing through machine learning. *Forest (C= 0.01, l2)*, **85**(15.78):1–39.

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., and Kim, J.-S., 2014. Analysis of off-target effects of crispr/cas-derived rna-guided endonucleases and nickases. *Genome research*, **24**(1):132–141.

Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B., *et al.*, 2018. Deepcrispr: optimized crispr guide rna design by deep learning. *Genome biology*, **19**:1–18.

Concordet, J.-P. and Haeussler, M., 2018. Crispor: intuitive guide selection for crispr/cas9 genome editing experiments and screens. *Nucleic acids research*, **46**(W1):W242–W245.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., *et al.*, 2016. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, **34**(2):184–191.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., and Root, D. E., *et al.*, 2014. Rational design of highly active sgrnas for crispr-cas9–mediated gene inactivation. *Nature biotechnology*, **32**(12):1262–1267.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J., 2019. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, **10**(1):390.

Gao, Y., Chuai, G., Yu, W., Qu, S., and Liu, Q., 2020. Data imbalance in crispr off-target prediction. *Briefings in bioinformatics*, **21**(4):1448–1454.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., *et al.*, 2022. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, **40**(2):163–166.

Hsu, P. D., Lander, E. S., and Zhang, F., 2014. Development and applications of crispr-cas9 for genome engineering. *Cell*, **157**(6):1262–1278.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., *et al.*, 2013. Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, **31**(9):827–832.

Kendall, A., Gal, Y., and Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H. R., Hwang, J., Kim, J.-I., and Kim, J.-S., 2015. Digenome-seq: genome-wide profiling of crispr-cas9 off-target effects in human cells. *Nature methods*, **12**(3):237–243.

Kingma, D. P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, .

Kuleshov, V., Fenner, N., and Ermon, S., 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR.

L Lun, A. T., Bach, K., and Marioni, J. C., 2016. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, **17**(1):1–14.

Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4):357–359.

Lazzarotto, C. R., Malinin, N. L., Li, Y., Zhang, R., Yang, Y., Lee, G., Cowley, E., He, Y., Lan, X., Jividen, K., *et al.*, 2020. Change-seq reveals genetic and epigenetic effects on crispr–cas9 genome-wide activity. *Nature biotechnology*, **38**(11):1317–1327.

Lin, J., Zhang, Z., Zhang, S., Chen, J., and Wong, K.-C., 2020. Crispr-net: A recurrent convolutional network quantifies crispr off-target activities with mismatches and indels. *Advanced science*, **7**(13):1903562.

Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., Gao, K., Hoang, L., Elibol, M., Doench, J. G., *et al.*, 2018. Prediction of off-target activities for the end-to-end design of crispr guide rnas. *Nature biomedical engineering*, **2**(1):38–47.

Liu, Q., Cheng, X., Liu, G., Li, B., and Liu, X., 2020. Deep learning improves the ability of sgrna off-target propensity prediction. *BMC bioinformatics*, **21**(1):1–15.

Liu, Q., He, D., and Xie, L., 2019. 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*, **15**(10):e1007480.

Mak, J. K., Störtz, F., and Minary, P., 2022. Comprehensive computational analysis of epigenetic descriptors affecting crispr-cas9 off-target activity. *BMC genomics*, **23**(1):805.

Manghwar, H., Li, B., Ding, X., Hussain, A., Lindsey, K., Zhang, X., and Jin, S., 2020. Crispr/cas systems in genome editing: methodologies and tools for sgrna design, off-target evaluation, and strategies to mitigate off-target effects. *Advanced science*, **7**(6):1902312.

Martin, F., Sánchez-Hernández, S., Gutiérrez-Guerrero, A., Pinedo-Gomez, J., and Benabdellah, K., 2016. Biased and unbiased methods for the detection of off-target cleavage by crispr/cas9: an overview. *International journal of molecular sciences*, **17**(9):1507.

Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., and Valen, E., 2014. Chopchop: a crispr/cas9 and talen web tool for genome editing. *Nucleic acids research*, **42**(W1):W401–W407.

Naeem, M., Majeed, S., Hoque, M. Z., and Ahmad, I., 2020. Latest developed strategies to minimize the off-target effects in crispr-cas-mediated genome editing. *Cells*, **9**(7):1608.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., *et al.*, 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, **32**.

Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., and Liu, D. R., 2013. High-throughput profiling of off-target dna cleavage reveals rna-programmed cas9 nuclease specificity. *Nature biotechnology*, **31**(9):839–843.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., *et al.*, 2014. Genome-scale crispr-cas9 knockout screening in human cells. *Science*, **343**(6166):84–87.

Sherkatghanad, Z., Abdar, M., Charlier, J., and Makarenkov, V., 2023. Using traditional machine learning and deep learning methods for on-and off-target prediction in crispr/cas9: a review. *Briefings in Bioinformatics*,

:bbad131.

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A., 2020. Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal*, **18**:2789–2798.

Soneson, C. and Robinson, M. D., 2018. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, **15**(4):255–261.

Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J., and Mateo, J. L., 2015. Cctop: an intuitive, flexible and reliable crispr/cas9 target prediction tool. *PloS one*, **10**(4):e0124633.

Störtz, F., Mak, J., and Minary, P., 2023. picrispr: Physically informed deep learning models for crispr/cas9 off-target cleavage prediction. *Artificial Intelligence in the Life Sciences*, :100075.

Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J., and Joung, J. K., 2017. Circle-seq: a highly sensitive in vitro screen for genome-wide crispr–cas9 nuclease off-targets. *Nature methods*, **14**(6):607–614.

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., *et al.*, 2015. Guide-seq enables genome-wide profiling of off-target cleavage by crispr-cas nucleases. *Nature biotechnology*, **33**(2):187–197.

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S., 2014. Genetic screens in human cells using the crispr-cas9 system. *Science*, **343**(6166):80–84.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., *et al.*, 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**:1–18.

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., *et al.*, 2015. Sequence determinants of improved crispr sgrna design. *Genome research*, **25**(8):1147–1157.

Yaish, O., Asif, M., and Orenstein, Y., 2022. A systematic evaluation of data processing and problem formulation of crispr off-target site prediction. *Briefings in Bioinformatics*, **23**(5):bbac157.

Zhang, S., Li, X., Lin, Q., and Wong, K.-C., 2019. Synergizing crispr/cas9 off-target predictions for ensemble insights and practical applications. *Bioinformatics*, **35**(7):1108–1115.

Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S., and Yang, S.-H., 2015. Off-target effects in crispr/cas9-mediated genome engineering. *Molecular Therapy-Nucleic Acids*, **4**:e264.

Zhang, Y., Long, Y., Yin, R., and Kwoh, C. K., 2020. Dl-crispr: a deep learning method for off-target activity prediction in crispr/cas9 with data augmentation. *IEEE Access*, **8**:76610–76617.

Zhu, L., Lei, J., Devlin, B., and Roeder, K., 2018. A unified statistical framework for single cell and bulk rna sequencing data. *The annals of applied statistics*, **12**(1):609.