Nucleic Acids Research, 2024, **52**, e87
https://doi.org/10.1093/nar/gkae759
Advance access publication date: 14 September 2024
Methods

Learning to quantify uncertainty in off-target activity for CRISPR guide RNAs

Furkan Özden * and Peter Minary *

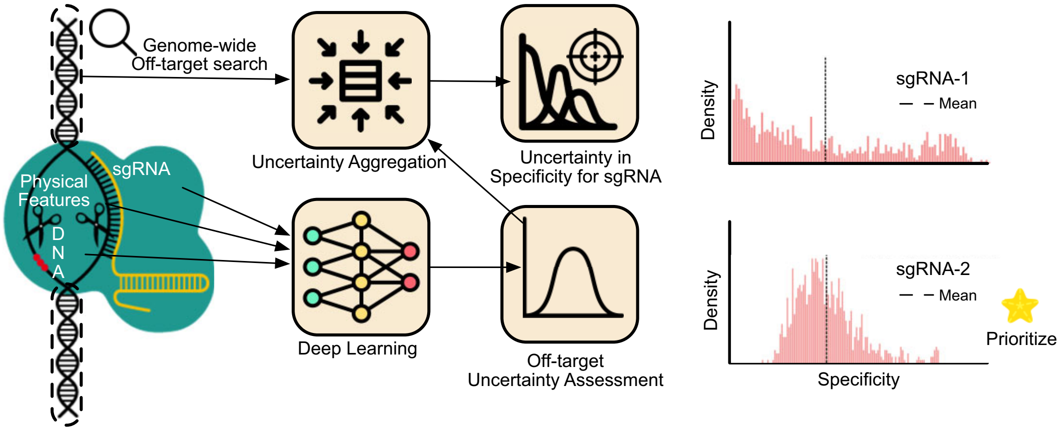
Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK

*To whom correspondence should be addressed. Tel: +44 7495704258; Email: furkan.ozden@cs.ox.ac.uk
Correspondence may also be addressed to Peter Minary. Email: 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 neural network architecture-based 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. 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. In addition, we present the first-of-its-kind genome-wide sgRNA efficiency score, crispAI-aggregate, enabling prioritization among sgRNAs with similar point aggregate predictions by providing richer information compared to existing aggregate scores. We show that uncertainty estimates of our approach are calibrated and its predictive performance is superior to the state-of-the-art *in silico* off-target cleavage activity prediction methods. The tool and the trained models are available at https://github.com/furkanozdenn/crispr-offtarget-uncertainty.

Graphical abstract



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 (1). CRISPR/Cas9 system quickly became among the most popular gene editing technologies recently with successful applications of editing eukaryotic genomes (2). CRISPR/Cas9 system has also been applied to knockout-screening studies (3), 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 the genome engineering field and has the potential to lead new treatments for genetic diseases (4).

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 the 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 the CRISPR/Cas9 system operates on the targeted DNA region, cleavage may also occur

Received: January 31, 2024. **Revised:** August 7, 2024. **Editorial Decision:** August 10, 2024. **Accepted:** August 23, 2024

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 (5). The existence of the off-target cleavage phenomenon has been one of the key factors limiting the development and applicability aspects of CRISPR-based genome editing systems. Studies found that few mismatch DNA sites are potentially recognizable by the sgRNA during the guiding process (6). Also, the off-target effects are dependent on many other factors, such as nucleosome occupancy, chromatin accessibility and both binding and heteroduplex energy parameters of the sgRNA of choice (7). However, many potential sgRNA sequences 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.

To date, many off-target cleavage activity prediction tools have been proposed to predict the potential off-target activity of a given sgRNA-target pair (i.e., targeted DNA sites and corresponding single guide RNA sequences). These algorithms use the data generated by experimental CRISPR off-target detection assays, such as GUIDE-seq (8), CHANGE-seq (9), DIGENOME-seq (10), CIRCLE-seq (11), and predict a point score for off-target cleavage activity for a given sgRNA-target 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 (12), logistic regression (13), gradient boosting (14) and ensemble learning (15) 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 (16). These models utilize novel sequence encoding strategies, feature engineering approaches by introducing physical features (17), class rebalancing techniques, and attention mechanisms (18) 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, MOFF score (19) stands out as a top-performing model that tackles the regression task of predicting the activity level of the sgRNA–target interface, which is the focus in this article as well. For a more detailed review of the current literature, Sherkatghanad *et al.* (20) presented a detailed overview of machine learning and deep learning-based studies for on/off-target activity prediction tasks 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 (20,21). This imbalance can make training routine machine learning models difficult, resulting in high accuracy 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 (14). 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, the only other study that incorporates uncertainty estimates into the off-target cleavage activity prediction task is Kirillov *et al.* (22), where the authors trained a Gaussian Process Regression model. 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 for 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, are often encountered in microbiome, bulk and single-cell RNA experiments (23,24). 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 (25). Addressing the highly imbalanced nature, and 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 (14). 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, and crispAI-aggregate to provide a probabilistic genome-wide specificity estimate for a given sgRNA.

Materials and methods

Datasets

For the training of crispAI, we used CHANGE-seq dataset (9). CHANGE-seq is a scalable, automatable tagmentation-based method for measuring the genome-wide activity of Cas9 *in vitro*. The authors identified 2 019 434 off-target sites on 110 sgRNAs across 13 therapeutically relevant loci in human primary T cells. Although CHANGE-seq assay is a high sensitivity *in vitro* genome-wide off-target detection method (20), it is still not able to detect all of the potential off-target sites due to the limited sensitivity of the experimental apparatus. However, studies suggested that off-target sites that have several mismatched positions with the respective sgRNA sequence (i.e. up to 6 bp) are *putative* off-target sites and are potentially harmful. We first scaled each read count so that the total reads for each sgRNA would be equal. Read counts are then linearly scaled to have a maximum read count of 10 000. In the dataset, 78 sgRNAs already fell within this range naturally. However, 32 sgRNAs initially had read counts exceeding the maximum count for their off-target sites, and among these, 5 sgRNAs showed particularly high read counts above 30 000 reads.

Many genome alignment-based methods (26–28) have been proposed for *in silico* discovery of the putative off-target sites for a given sgRNA. We employed one of the most popular, light-weight tool CasOFFinder (27) 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 allowed base-pair mismatched positions with the sgRNA sequence. We obtained a total of 1 78 801 putative off-target sites, yielding a total of 1 581 757 off-target sites that are not also in CHANGE-seq dataset.

To evaluate crispAI, we used five different test sets obtained with different assays. More specifically: (i) we randomly splitted 10% of the CHANGE-seq dataset on human primary T-cells obtaining 168 465 off-target sites for 110 sgRNAs; (ii) GUIDE-seq dataset containing 443 off-target sites for 13 sgRNAs; (iii) SITE-seq dataset containing 6,097 off-target sites for 8 sgRNAs; and (v) HEK293T K562 cell-lines datasets used in Chuai *et al.* (29) containing 536 and 120 off-target sites for 12 and 18 sgRNAs respectively. These datasets represent a diverse range of off-target detection methods. GUIDE-seq is a global detection method for DNA double-stranded breaks introduced by CRISPR RNA-guided nucleases. It works by capturing double-stranded oligodeoxynucleotides to identify off-target cleavage activities. This approach provides a comprehensive view of potential off-target sites across the genome. SITE-Seq, on the other hand, is a biochemical method specifically designed to identify off-target cleavage sites of CRISPR-Cas9 RNA-guided endonucleases. It employs a technique of selective enrichment and sequencing of adapter-tagged DNA ends to pinpoint off-target locations. Both these methods complement the CHANGE-seq data, offering additional perspectives on off-target detection.

Modelling of crispAI

Problem formulation

Letting $x_s \in \{0, 1\}^{m_s \times \ell}$, where m_s and ℓ are feature dimension and sequence length respectively, be the nucleotide sequence-based features of sgRNA-target pair interface, $x_p \in [0, 1]^{m_p \times \ell}$, where m_p is feature dimension for physical features, denote the physical descriptor based features of sgRNA–target pair interface and $y \in \mathbb{R}$ denote the cleavage read depth, we model the 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:

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

$$f_{\mathrm{w}} : (\{0, 1\}^{m_s \times \ell}, [0, 1]^{m_p \times \ell}) \rightarrow \mathbb{R}^3, x \rightarrow (\pi, \mu, \theta) \tag{2}$$

where ZINB(k; π, μ, θ) is the ZINB distribution with parameters π, μ and θ for the mixture coefficient representing the point-mass (δ_0) at 0, mean and dispersion of the negative binomial component (NB), respectively. We model the parameters for the conditional distribution with a multi-input, multi-output parametric function $f_{\mathrm{w}} : (\{0, 1\}^{m \times \ell}, [0, 1]^{m_p \times \ell}) \rightarrow \mathbb{R}^3$ with parameter set w. Note that the output space of the function f_{w} is $\mathbb{R}^3$, where three output dimensions represent π, μ and θ. Thus, using a data-set consisting of N samples in the form of 3-tuples, $\mathcal{D} = \{(x_s^i, x_p^i, y^i)\}_i^N$, we want to be able to calculate the posterior distribution of the cleavage score y given the features x_s, x_p and the data $\mathcal{D}$ — that is $\mathbb{P}(y|x_s, x_p, \mathcal{D})$.

Encoding of the sequence-based sgRNA-target interface features

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 × 23 binary matrices for each sequence. Then we employed the sgRNA-target pair encoding scheme proposed in Lin et al. (30). We represent any sgRNA-target sequence pair with a 6 × 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, the resulting 4 × 23 binary matrix shows the mismatches between the 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 '$T \rightarrow T$', obtaining a 6 × 23 matrix for the sequence-based features of the sgRNA-target pairs.

Physical descriptors and encoding of the 147-bp sequence context of the target site

We used 147-bp sequence context on the off-target loci (i.e., 73-bp flank on each side of an off-target sequence position) with a sliding window approach to obtain: (i) Nucleotide BDM score (31); (ii) GC content; (iii) NuPoP occupancy and (iv) NuPoP affinity scores (32) and normalized the obtained scores in the [0,1] interval. Hence, we obtained 4 × 23 matrix for physical descriptors of the off-target sequence context. GC count refers to the proportion of G and C bases within the 147-bp sliding window, centered around a given off-target sequence position. Nucleotide BDM is a training-free method to approximate the algorithmic complexity of a given DNA sequence. NuPoP scores refer to a Hidden Markov Model (HMM), trained to estimate the nucleosome affinity and occupancy at single base-pair resolution. For a more detailed discussion on the effects of these features to CRISPR-based cleavage activity, we refer to Störtz *et al.* (17). Additionally, we trained a sequence-only version of crispAI for comparison with the full model as an ablation study. We validated that, physical descriptors improved performance (Supplementary Table S11).

Architecture and training of crispAI

crispAI is an end-to-end multi-output CNN and bi-LSTM fusion neural network specifically designed to quantify the uncertainty in off-target cleavage activity of sgRNA-target pairs for CRISPR/Cas9 system. We show that crispAI architecture can increase performance on off-target cleavage activity prediction task on different test sets, while providing uncertainty quantification in off-target activity (Results). We conducted comprehensive evaluations of different architectural configurations to identify the most effective model. Spearman Correlation coefficients for seven predictive models, including five different models trained with samples from the searched hyperparameter space and CNN-only and biLSTM-only versions of the original architecture, across the CHANGE-Seq, GUIDE-Seq, and SITE-Seq, HEK293T and K562 cell-lines test datasets are presented in Supplementary Tables S8 and S9. These configurations range from simpler models, which feature the smallest CNN layers and LSTM, to more complex ones with the largest CNN and LSTM layers. The performance metrics indicate that the best model achieved the highest correlation for the CHANGE-Seq dataset (0.5114), while others performed best for GUIDE-Seq and SITE-Seq respectively. The model is optimized on the validation split of CHANGE-seq, and none of the model optimizations have seen either CHANGE-seq or other test sets.

crispAI architecture is designed to estimate the parameters of the ZINB distribution conditioned on the sgRNA-target 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 three 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 three associated parameters for the ZINB distribution, π, μ and θ, are discussed below. The architecture can be formulated through the following equations:

$$
\begin{aligned}
E_{cnn} &= \mathrm{ReLU}(\mathrm{CNN}(X_s)_{i=1}^{3}),\\
E_s &= \mathrm{ReLU}(\mathrm{BiLSTM}(E_{cnn})),\\
E_p &= \mathrm{ReLu}(\mathrm{CNN}(X_p))\\
E_{total} &= \mathrm{Concat}(\mathrm{ReLU}(\mathrm{FC}_1(E_s)), \mathrm{ReLU}(\mathrm{FC}_2(E_p))),\\
\pi &= \mathrm{Logit}(\mathrm{FC}_\pi(\mathrm{ReLU}((\mathrm{FC}_3(E_{total}))))),\\
\mu &= \mathrm{Exp}(\mathrm{FC}_\mu(\mathrm{ReLU}((\mathrm{FC}_3(E_{total}))))),\\
\theta &= \mathrm{Exp}(\mathrm{FC}_\theta(\mathrm{ReLU}((\mathrm{FC}_3(E_{total})))))
\end{aligned}
\tag{3}
$$

where E_{cnn}, E_s, Ep and E_{total} represent CNN encoding of the sequence-based features, biLSTM encoding of the CNN extracted features, CNN encoding of physical descriptors and concatenated total encoding of the sequence-based and physical descriptor features respectively. To ensure positivity on parameters π and μ, we use exponential activations, and for drop-out parameter π, we use Logit activation 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 ($\mathcal{L}_{\mathrm{ZINB}} = \mathrm{NLL}_{\mathrm{ZINB}}$) for three parameters of the ZINB distribution.

$$
\mathcal{L}_{\mathrm{ZINB}}(\pi, \mu, \theta, y) = -\textstyle\sum_{i=1}^{N} \log\left(\pi \delta_0(y_i) + (1-\pi)\mathrm{NB}(y_i; \mu, \theta)\right), \tag{4}
$$

where all parameters in (4) are defined in the Problem Formulation section. We performed ablation studies with candidate zero-inflated and non-hurdle versions of the distributions, including Zero-Inflated Poisson (ZIP), Negative Binomial (NB), Poisson, and Zero-Inflated Negative Binomial (ZINB). Supplementary Table S6 shows that the model with the ZINB module outperformed others across most datasets, except the K562 cell line, where the Negative Binomial module achieved the highest Spearman Correlation coefficient. Furthermore, another ablation study detailed in Supplementary Table S7 demonstrated the goodness of fit for these distributions when applied to the CHANGE-seq test dataset.

We splitted CHANGE-seq dataset into training, validation and testing sets containing 70%, 20% and 10% of the samples respectively. For all other test datasets, we left the data as is and did not augment the samples. We used Adam optimizer (33) 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 (34). Finally, we used SCVI-tools (35) to implement the loss functions. The model is trained on a single 24G NVIDIA TITAN RTX GPU. The training duration for the model on CHANGE-seq datasets is around ∼50 minutes on the said GPU.

Genome-wide sgRNA specificity prediction with crispAI-aggregate score

We designed, the first of its kind, crispAI-aggregate score for uncertainty-aware sgRNA genome-wide specificity prediction, inspired by MOFF-aggregate score (19). First, we use CasOFFinder (27) to search putative off-target DNA sites up-to a user-specified (e.g. $N = 5$) number of mismatches genome-wide for the sgRNA of interest. Then, we obtain crispAI-predicted cleavage activity distributions for each obtained putative off-target site. Having obtained the posterior distributions, we first sample each predicted distribution n_{samp} times, where n_{samp} is the number we wish to sample crispAI-aggregate distribution (Supplementary Note 1). We then, calculate the ratio between the summation of the crispAI-predicted distribution samples of all of the detected off-target sites and the crispAI-predicted distribution samples for the on-target site. Finally, crispAI-aggregate score for a given sgRNA is defined as the log of the obtained ratio.

$$
\mathbb{P}(y_{sg}|X_{sg}, M) = \log\left(\frac{\sum_{i=0}^{M} \mathbb{P}(y^i|x_s^i, x_p^i)}{\mathbb{P}(y^{\mathrm{on}}|x_s^{\mathrm{on}}, x_p^{\mathrm{on}})}\right), \tag{5}
$$

where y_{sg} is the genome-wide specificity score, X_{sg} represents the sequence-based features of the sgRNA of interest, M is the number of detected off-target sites, which depends the user-specified maximum number of mismatches, N (N is one of the inputs of the genome-wide search algorithm, which yields M samples.) The variables on the right hand side of (5) are defined in (1) and indexed by the off-target site indice in the nominator and the on-target site indice in the denominator.

Comparison and Implementation Details

We implemented crispAI using Pytorch (version 2.1.0), leveraging its robust framework for building and training deep learning models. Additionally, we utilized Pytorch distributions and Numpyro (version 0.12.1) to manage and implement various probability distributions essential for our model's uncertainty quantification. Also we employed scvi-tools for implementing loss functions, which provided a comprehensive suite for our model training and evaluation needs.

We benchmarked crispAI against several models spanning heuristic-based approaches, traditional machine learning, and deep learning techniques. The models used for comparison were: MIT Score, CFD Score, MOFF, CRISPR-Net, CNNCrispr and CRISOT score and Elevation-aggregate, CRISPRoff-spec, MOFF-aggregate, CNN_std and CRISOT-spec for the aggregation task.

The CFD score evaluates the off-target propensity of sgRNA-DNA interactions by assigning scores based on the location and type of mismatches, with the final score being the product of individual mismatch scores. We used source code of CFD score calculator provided by John Doench to iGWOS (36) https://github.com/bm2-lab/iGWOS/blob/master/CFD/otscore.py. The MIT scoring model, introduced by Hsu *et al.*, assesses CRISPR off-target effects by evaluating different mismatch positions and using a weight matrix to determine off-target efficiency. Again we used the implementation available at iGWOS repository https://github.com/bm2-lab/iGWOS/blob/master/MIT/otscore.py. CNNCrispr is a deep learning-based model that integrates the GloVe model for feature representation and combines RNN with CNN to predict off-target propensity using sequence information alone. It avoids biases from manual feature construction and demonstrates superior prediction abilities. We obtained CNNCrispr source code and trained models from https://github.com/LQYoLH/CnnCrispr. CRISPR-Net, another complex deep learning model, uses an LRCN-based neural network with an Inception-based convolutional layer and bi-directional LSTM units to preserve spatial information of the sequence pair, enhancing its predictive performance. We obtained CRISPR-Net source code and trained models from https://github.com/JasonLinjc/CRISPR_Net. MOFF is a statistical model-based off-target predictor that combines individual mismatch effect, combinatorial effect and GMT effect, outperforming more complex deep learning frameworks in predicting off-target effects. We followed the installation guide at https://github.com/MDhewei/MOFF and used MOFF command-line program for comparison. CRISOT, a comprehensive suite for off-target prediction and sgRNA optimization, leverages molecular dynamics simulations and AI to analyze RNA-DNA interactions, using the XGB algorithm based on tree boosting and logistic regression to provide a robust and scalable platform for sgRNA design and reducing off-target effects in various CRISPR systems. We obtained trained CRISOT models and source code from https://github.com/bm2-lab/CRISOT and used respective commands for benchmarking of each task.

We compared crispAI against five baseline models. Two of these models are simpler machine learning approaches: quantile regression and random forest. The remaining three models are extensions of crispAI, incorporating Poisson, and Negative Binomial (NB) distributions, which are equally complex in terms of model architectures and parameter counts. We implemented quantile regression using the QuantileRegressor from the sklearn package. The model was trained for each quantile with a step size of 0.005 up to 1, using an alpha value of 1, intercept set to true, and 'highs' as the solver. For the random forest model, we used the RandomForestRegressor from the sklearn.ensemble package, configuring it with 1000 trees (n_estimators = 1000), the squared error criterion (criterion = 'squared_error') and a minimum samples split of 2 (min_samples_split = 2). For the extended versions of crispAI, which included the ZIP, Poisson and NB models, we maintained the same optimized hyperparameter configuration as the original crispAI model. The key differences in these models are the parameters they predict: the ZIP model predicts two parameters (π and λ), the Poisson model predicts one parameter (λ), and the NB model predicts two parameters (θ and μ). We have analyzed the running time of crispAI in various operations. Specifically, we evaluated: (i) crispAI off-target scoring inference running time, (ii) crispAI-aggregate score versus the number of allowed mismatches between the sgRNA and target sequence and (iii) sampling time from the obtained conditional posterior distributions for both crispAI-aggregate and off-target scores. The running time experiments were conducted on a single NVIDIA TITAN RTX GPU (24 GB, 384-bit). For the off-target scoring inference, the model processed 131 712 samples in approximately 578 s with a batch size of 128, which is the models default batch size for the command line interface. For crispAI-aggregate score calculations, the running times ranged from 13.94 seconds for two mismatches to 31.26 s for 9 mismatches, sampling each obtained specificity distribution 1000 times. For posterior predictive sampling time, the model took 0.016, 0.059, 0.231, 2.197, 4.465 and 6.567 s for generating 100, 1000, 10 000, 100 000, 200 000 and 300 000 samples, respectively.

Results

Overview of crispAI

We designed crispAI to be a hybrid deep learning architecture based on the count distribution Zero-Inflated Negative Binomial (ZINB), which accounts for the 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 Convolutional Neural Network (CNN) and bi-directional Long Short Term Memory (biLSTM) layers to extract sequence-based features of the sgRNA-target pair, which are encoded using a binary matrix encoding scheme first presented by Lin et al. (30). We used an additional CNN layer to process physical descriptors of the sequence context, namely: (i) Block Decomposition Method (BDM) score (31); (ii) GC content; (iii) NuPoP Occupancy and (iv) NuPoP Affinity scores (32). The importance of these descriptors for off-target cleavage activity has been highlighted by a recent study (7). Both sequence-based features and physical descriptor features are used to extract features related to the off-target cleavage activity of the sgRNA–target interface. The extracted features are then concatenated in a late fusion fashion for the final Fully Connected (FC) layer to predict three parameters, μ, θ and σ, associated with the ZINB distribution. We trained the network weights based on a loss function optimising the likelihood of the observed data (Figure 1).

Following this design, we experimented to demonstrate the advantage of incorporating physical descriptors in

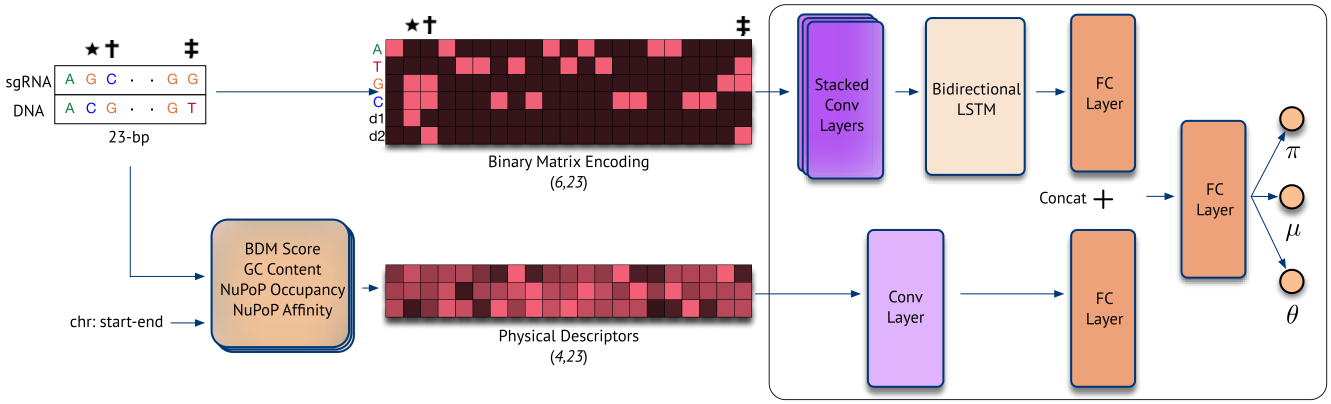


Figure 1. Workflow figure. Logical OR operation is employed to encode the sgRNA-target interface to a matrix of shape (6,23) where the first four rows represent sequence letters (i.e. ATGC) and the last two rows represent mismatch direction. Star, cross and double-cross signs represent mismatched positions and their corresponding encoding. Physical descriptors of the sequence context such as (i) BDM Score; (ii) GC content; (iii) NuPoP occupancy and (iv) NuPoP affinity are calculated and normalized values are encoded into a matrix of shape (4,23) for each sequence position. Then, the sequence encoding features are extracted with a series of Convolutional (Conv) layers followed by a Bi-LSTM layer and features of physical descriptors are extracted with a Conv layer. Both extracted feature vectors are mapped to 128 − *d* vector spaces with fully connected (FC) layers and are concatenated. The concatenated features are then passed to a final FC layer to predict three parameters π, μ and θ.

our model. This analysis, presented in Supplementary Figure S1, illustrates crispAI's ability to differentiate between identical off-target sequences located in different genomic regions. We identified two regions with identical sgRNA and target sequences: one in a noncoding region (chrX:64118237-64118260:–) and another in the LINC00374 gene (chr13:58228371:58228394:+). Despite having the same sequence, these regions exhibit different physical properties due to their distinct genomic locations. Supplementary Figure S1A shows the conditional posterior distributions obtained using crispAI for both samples, along with their corresponding physical descriptors. The mean crispAI-scores were 0.0134 and 0.0181, matching the magnitude relation of the ground-truth detected activity scores of 0.0054 and 0.0288 for the non-coding and LINC00374 regions, respectively. Supplementary Figure S1B further illustrates how physical features such as GC content, Nucleotide BDM, NuPoP affinity and occupancy scores enable the model to distinguish between these otherwise identical off-target sites. This experiment highlights the importance of physical descriptors in improving model performance and providing more nuanced predictions for off-target sites. It is worth noting that, a model using only sequence-based features would not be able to differentiate these samples.

Enabling uncertainty quantification for off-target activity prediction

Our proposed architecture, crispAI, models the off-target cleavage activity in a probabilistic framework. Hence, making it possible to sample from the posterior off-target activity distributions conditioned on both the sequence-based and physical features of the sgRNA-target interface. Figure 2 depicts example distributions for eight randomly selected test samples from the left-out test portion of the CHANGE-seq (9) dataset, where point predictions and the ground truth CHANGE-seq detected off-target cleavage activity values are shown with vertical lines.

Additionally, we expanded our analysis to a broader set of sgRNA–target interfaces, as illustrated in Supplementary Figure S2. The figure presents a comprehensive comparison of crispAI's predictive capabilities against seven other widely-used methods: CNNCrispr, CFD-score, MIT-score, MOFF, CRISPR-Net, CRISOT (37) and CRISPRoff (38). For 16 randomly sampled sgRNA–target pairs, we plotted crispAI's posterior distributions alongside point predictions from these competing methods. This extended analysis reveals significant heteroscedasticity in predictions across models, highlighting the inherent challenges in accurately estimating off-target activity. Notably, crispAI's probabilistic framework demonstrates several advantages. For low mismatch count interfaces, where many methods tend to predict uniformly high values, crispAI provides more nuanced distributions that often align better with the ground truth CHANGE-seq data. This is particularly evident in cases such as those shown in row 2 column 4 and row 3 column 3 of Supplementary Figure S2, where crispAI's distributions capture the lower actual activity levels missed by some other models. This extended analysis not only reinforces the capabilities of crispAI in modeling off-target activity but also highlights its robustness across a diverse range of sgRNA–target interfaces. By providing full posterior distributions rather than single point estimates, crispAI offers a more comprehensive view of potential off-target effects. Furthermore, for off-target sites with similar ground truth activity values, crispAI generates distinct posterior distributions with comparable expected values, matching the observed data. This approach contrasts sharply with the highly variable point predictions of other models for such cases. For instance, the subplots at row 2 column 3, row 1 column 1 and row 4 column 1 in Supplementary Figure S2 demonstrate how crispAI consistently captures the similarity in ground truth values while other models show significant inconsistencies in their predictions.

We started evaluating crispAI by visualising the uncertainty estimates for individual sgRNA-target pairs on the left-out test portion of the CHANGE-seq dataset (9). CHANGE-seq is a highly scalable, 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. We randomly splitted 10% of the CHANGE-seq dataset for testing obtaining 168 465 samples in total. First,

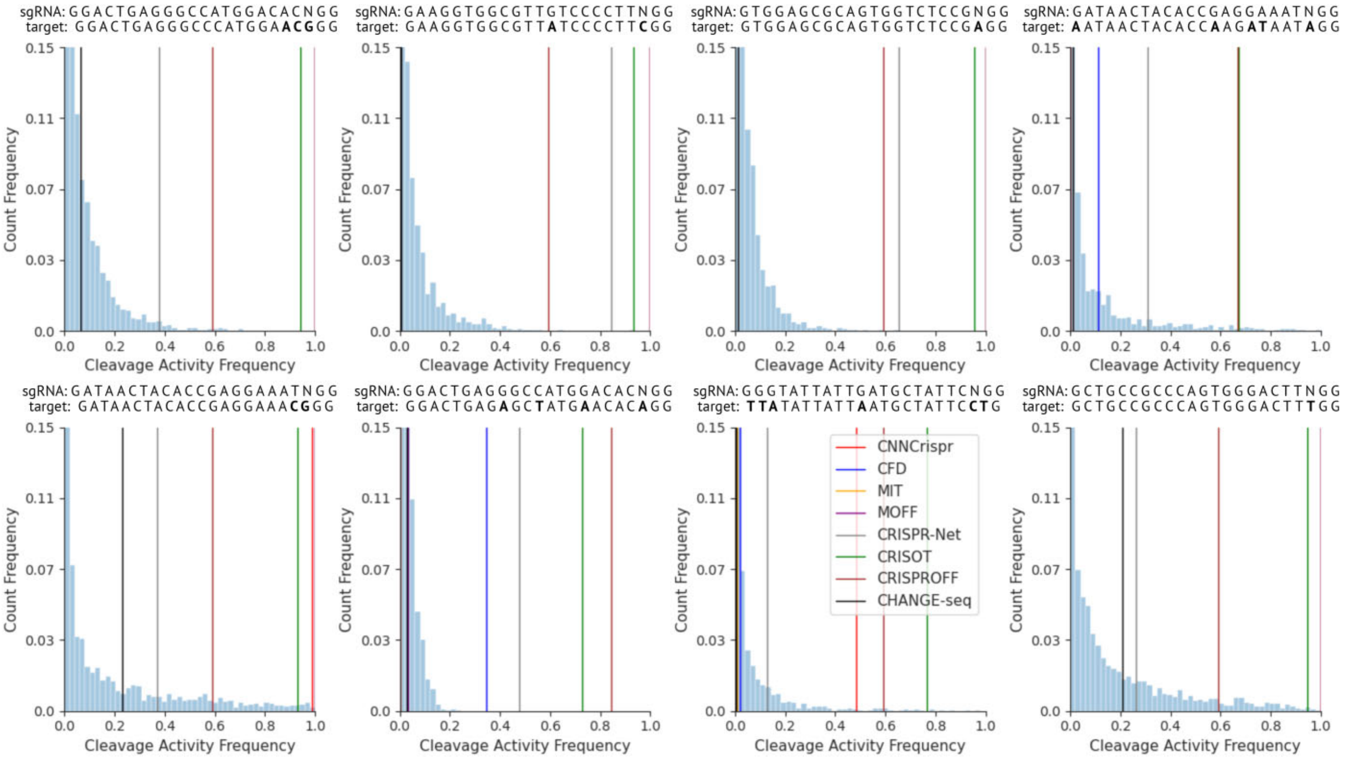


Figure 2. Example distributions for eight randomly selected sgRNA-target interfaces. Our approach, crispAI, enables sampling from the posterior off-target activity distribution conditioned on both sequence-based and physical descriptor-based features of the sgRNA-target pair of interest. Vertical lines represent CnnCrispr, CFD, MIT, MOFF, CRISPR-Net, CRISOT, CRISPRoff scores and ground truth CHANGE-seq detected off-target cleavage activity frequencies for each individual sgRNA-target pair. The title of each plot shows the sgRNA sequence and its corresponding off-target sequence, with bold characters highlighting mismatches between them. The histograms show the predicted cleavage activity distributions for each sgRNA–target pair, with the x-axis representing cleavage activity frequency and the y-axis showing probability. The plots demonstrate how crispAI can provide probabilistic predictions of off-target activity, allowing for uncertainty quantification in contrast to single-point estimates from other methods.

we obtained posterior parameters of the respective ZINB distributions conditioned on the sgRNA-target pair features for each sample using crispAI on the left-out test portion of the CHANGE-seq dataset. We sampled corresponding posterior distributions many times to obtain empirical Probability Mass Functions (PMFs) for each sample. Then, we obtained CnnCrispr score (16), CFD score (12), MIT score (39), MOFF score (19), CRISPR-Net score (30), CRISOT score (37) and CRISPRoff score (38) on the same test portion for comparison with the ground truth CHANGE-seq detected activity values and expected value of the crispAI-predicted PMFs. We sorted the test samples and all associated point predictions along with the ground truth values based on their predicted Upper Confidence Bound values (UCB). Figure 3A shows the confidence intervals of crispAI-predicted posterior PMFs, point predictions of competing methods and the expected value of the predicted PMFs and the ground truth CHANGE-seq detected activity values. We observed that predicted intervals accurately captured the ground truth cleavage activity values for almost all of the samples. For on-target sites and highly active off-target sites (e.g., activity frequency > 0.25) the posterior distributions yielded very high 95% CI UCBs reaching up to maximum activity frequency. This 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 (9). Additionally, we observed that the expected values of crispAI-predicted PMFs closely resembled the ground truth activity values, whereas CnnCrispr score, CFD score, MIT score and CRISPR-Net scores are generally far off while MOFF score is somewhat better at distinguishing between minimal activity and highly active sites (Supplementary Table S2). Due to the vast imbalance between minimal activity off-target sites (e.g. <0.07) and higher activity off-target sites we splitted the comparison into two parts. Figure 3B shows the samples with ground truth CHANGE-seq detected activity values >0.07 for better visibility. While MOFF, CRISOT-FP and MIT scores were almost always in the 95% CI for the minimal activity off-target sites, the other competing methods overestimated the ground-truth activity value. Whereas for higher activity off-target sites for only half of the test samples, MOFF and MIT were in the 95% CI, again, other methods generally overestimated the ground-truth activity value. We observed that the expected values of the crispAI-predicted conditional posterior distributions for respective off-target sites closely resemble the CHANGE-Seq detected activity values for both minimal and highly active off-target sites.

To further emphasize the imbalance in high activity and minimal activity off-target sites, we analyzed the test portion of the CHANGE-seq dataset. We categorized the off-target sites into two distinct groups based on their read counts: those with read counts <50 and those with read counts ≥50. For the first group, we divided the off-target sites into 10 bins, each representing a range of read counts. We observed that the majority of off-target sites fell into the lowest activity bin, with

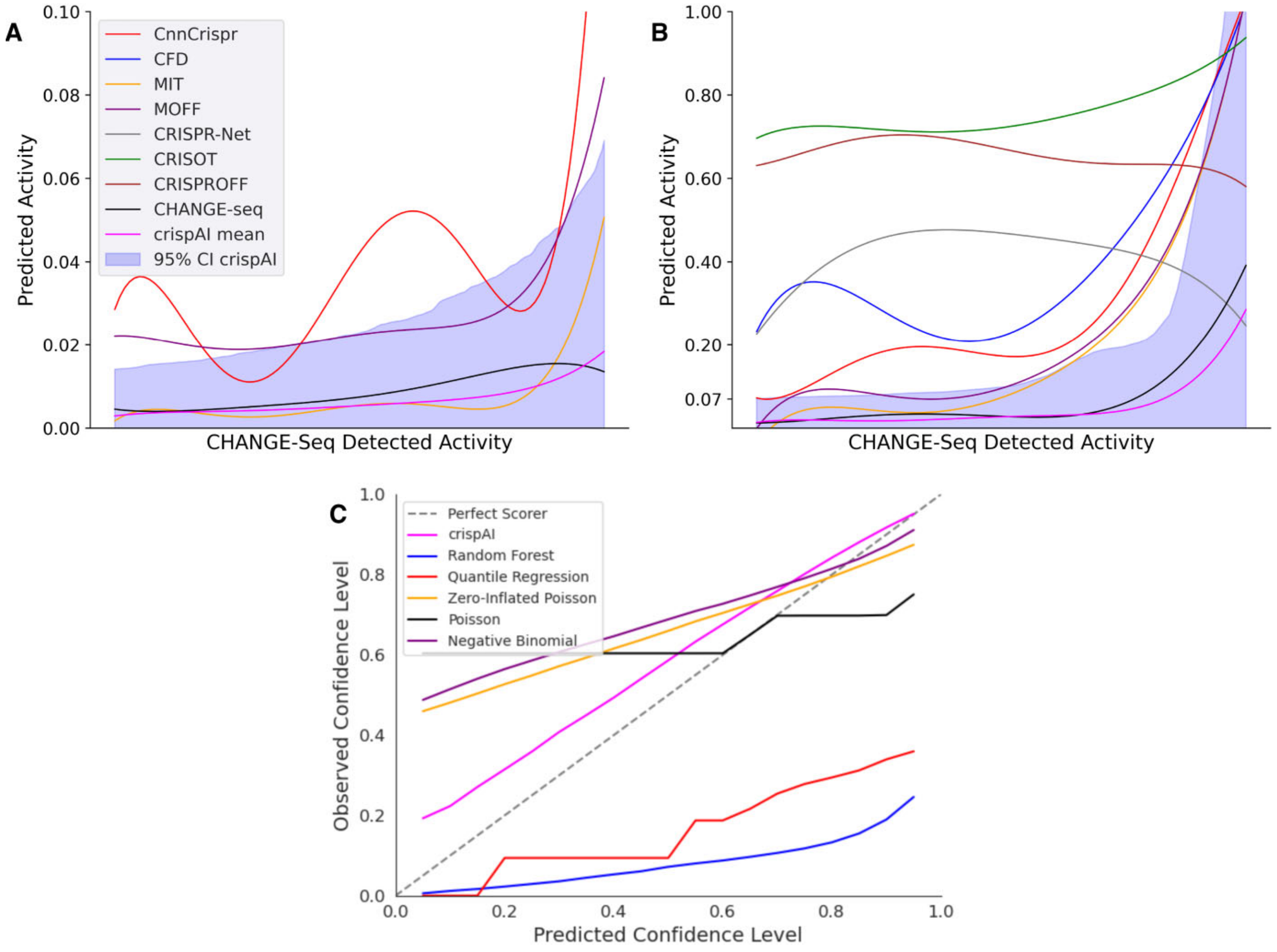


Figure 3. Depiction of crispAI-predicted uncertainty estimates on the test portion of the CHANGE-seq data. **(A)** First, we obtained the crispAI-predicted off-target activity distributions for samples with positive CHANGE-seq counts. Then we calculated both the 95% confidence interval and the expected value (green) for each distribution and sorted the off-target samples and associated predictions of CnnCrispr score (black), CFD score (blue), MIT score (red), MOFF score (purple), CRISPR-Net score (gray), CRISOT (brown), CRISPRoff (pink) and mean of crispAI-obtained posterior predictive distribution (green) along with the ground truth CHANGE-seq (red) detected activity (normalized between 0 and 1) in the increasing order, based-on the predicted Upper Confidence Bound (UCB) values. Due to vast imbalance of low-activity samples (e.g. $y < 0.07$) the confidence interval plot is splitted to two parts. **(B)** Part (A) is repeated for cleavage activity values greater than 0.07. **(C)** Observed versus Predicted confidence level plot is depicted as the uncertainty diagnostics plot. For benchmarking purposes, we generated baseline uncertainty estimates by training two smaller models: i) Random Forest and ii) Quantile Regression models; and three larger models, equally complex with crispAI using different distributions: Zero-Inflated Poisson (ZIP), Poisson and Negative Binomial (NB). All models are trained on the same training portion of the CHANGE-seq dataset.

read counts between 1.0 and 5.8, which contained 13 609 samples, or 82.03% of the total samples in this group. This stark contrast highlights the significant skew towards low-activity off-target sites. As the read counts increased, the number of off-target sites decreased dramatically. For example, Bin-2 (with read counts between 5.8 and 10.6) contained 1147 samples (6.91%), while Bin-10 (with read counts between 44.2 and 49.0) contained only 36 samples, representing a mere 0.22% of the total. We ran the crispAI-score on each bin and calculated the average coefficient of variation of the obtained distributions and the average crispAI point prediction score. The mean crispAI scores closely matched the mean CHANGE-seq read counts across the bins, demonstrating the model's accuracy in predicting minimal activity sites. However, the coefficient of variation increased for higher activity bins, indicating greater uncertainty in predicting these sites. For the second group, representing off-target sites with read counts of 50 or higher, we again divided the data into bins, each containing a range of read counts. This analysis showed an even more pronounced imbalance. Bin-11, with read counts between 50.0 and 529.4, contained 441 samples (2.66% of the total), while the highest activity bins (Bins 12 through 20) had extremely low sample counts, with most bins containing only 1 to 4 samples. Some bins, like Bin-18, had no samples at all. This scarcity of high-activity off-target sites is evident in the ratios, which range from 0.0011 to 0.0001, representing less than 0.1% of the total samples for each of these high-activity bins. Due to the extreme rarity of these high-activity sites, there was more variability in the predictions. For some very high activity bins (e.g., Bin-14 and Bin-16), crispAI's mean reads were significantly lower than the CHANGE-seq reads, due to the model's conservative predictions for these rare events. This analysis, presented in Supplementary Tables S13 and S14, underscores the challenges posed by the imbalance between minimal and higher activity off-target sites. It highlights the importance of developing robust models that can effectively handle such imbalances, ensuring accurate prediction and assessment of off-target effects in CRISPR-based genome editing.

Ideally, in single 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.

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.* (40). 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 comparison with baseline uncertainty estimates, we trained two simpler and three complex models: (i) Random Forest regressor; (ii) Quantile Regression; (iii) Zero-Inflated Poisson (ZIP); (iv) Poisson (v) and Negative Binomial (NB) models on the same data crispAI is trained. We trained the latter three models using a modified version of crispAI architecture, where we modified the output layer of the model to handle each respective probability distribution parameters. Hence, the latter three models are equally large and complex with the original crispAI model. Figure 3C depicts diagnostic lines for each method. We observed that diagnostic plot of crispAI-predicted CI levels is similar to the ideal calibration line, while baseline uncertainty estimates deviated either by higher underestimation or over-estimation.

Improving *in silico* CRISPR/Cas9 off-target cleavage activity prediction performance with crispAI

To evaluate the predictive performance of crispAI, we used five test datasets: (i) left-out test portion of the CHANGE-seq dataset (9); (ii) GUIDE-seq dataset (8); (iii) SITE-seq dataset (41) and the datasets presented in Chuai *et al.* (29) (iv) HEK293T cell-line and (v) K562 cell-line (Datasets). CHANGE-seq is a scalable, automatable tagmentation-based method for measuring the genome-wide activity of Cas9 *in vitro*. GUIDE-seq is a method for globally detecting DNA double-stranded breaks introduced by CRISPR RNA-guided nucleases (RGNs), using the capture of double-stranded oligodeoxynucleotides to identify off-target cleavage activities. SITE-Seq is a biochemical method for identifying off-target cleavage sites of CRISPR-Cas9 RNA-guided endonucleases through the selective enrichment and sequencing of adapter-tagged DNA ends. We inputted sgRNA sequence, target sequence and associated coordinates of each sgRNA-target pair in the mentioned datasets to crispAI pipeline, and obtained crispAI-predicted posterior off-target activity score distributions as depicted in Figure 4A.

We observed that point predictions obtained using the expected values of crispAI-predicted distributions significantly out-performed all of the competing tools with respect to Spearman correlation coefficient on all test datasets except SITE-seq. It is worth noting that crispAI is not re-trained for any of the test sets described. crispAI-predicted distributions performed second-best following CnnCrispr for this dataset. Specifically we observed, 21.09%, 6.92% and 17.77% improvements over the best models in HEK293T, K562 and union of the two test datasets respectively (Figure 4B). Similarly for the CHANGE-seq and the GUIDE-seq datasets we observed, 19.51%, 10.76% improvements over best performing tools MOFF and CRISPR-Net respectively and 12.01% deterioration on SITE-seq dataset performing second best behind best performing tool CnnCrispr (Supplementary Table S1). Also, we stratified samples in CHANGE-seq test dataset to two folds as minimal and high activity and measured performance of each competing model. For high activity sites, crispAI achieves the highest Spearman correlation of 0.4529, significantly outperforming other models. The next best performers are MIT and CRISPR-Net, both with a correlation of 0.2634. For minimal activity sites, crispAI again leads with a correlation of 0.1030, followed closely by CRISPR-Net at 0.1029. In both test folds, crispAI performs best, with a notably higher correlation in the high activity portion (Supplementary Table S12).

Additionally, we plotted 5 box-plots, one for each dataset, illustrating the distribution of Coefficient of Variation of crispAI-predicted distributions to compare with the variances of the ground truth cleavage activity values given in the datasets. We observed increased median lines and third quartile lines as the variance of the dataset increases as expected. For the sake of obtaining standard deviations of performance measure on CHANGE-seq test set, we performed a 10-fold cross validation on each test dataset, using the same trained model for each fold (Supplementary Table S10).

Effects of mismatches and off-target activity on the uncertainty of the predictions.

We investigated whether crispAI-predicted off-target activity distributions captured effects of number of mismatched positions in sgRNA-target pair and the ground truth assay detected off-target activity values for the CRISPR/Cas9 system. To achieve this, we used the left-out test portion of CHANGE-seq dataset and stratified the samples with respect to: (i) the base-pair mismatch count between the sgRNA and target sequences and (ii) detected CHANGE-seq frequency. We then plotted the stratified folds with respect to the coefficient of variation of the associated crispAI-predicted off-target activity distributions.

We observed a consistent decrease in the coefficient of variation span as the mismatch count decreased from 6 to 0. (Figure 5B). This result is expected, as the number of allowed mismatches between sgRNA and target sequences increases, allowed degree of freedom for other sequence-based factors (e.g. that are shown to be correlated with higher off-target cleavage activity) increases. These factors include but not limited to GC content, mismatch location and specific ordering of nucleotides. This yields a higher coefficient of variation in predicted off-target activity distributions.

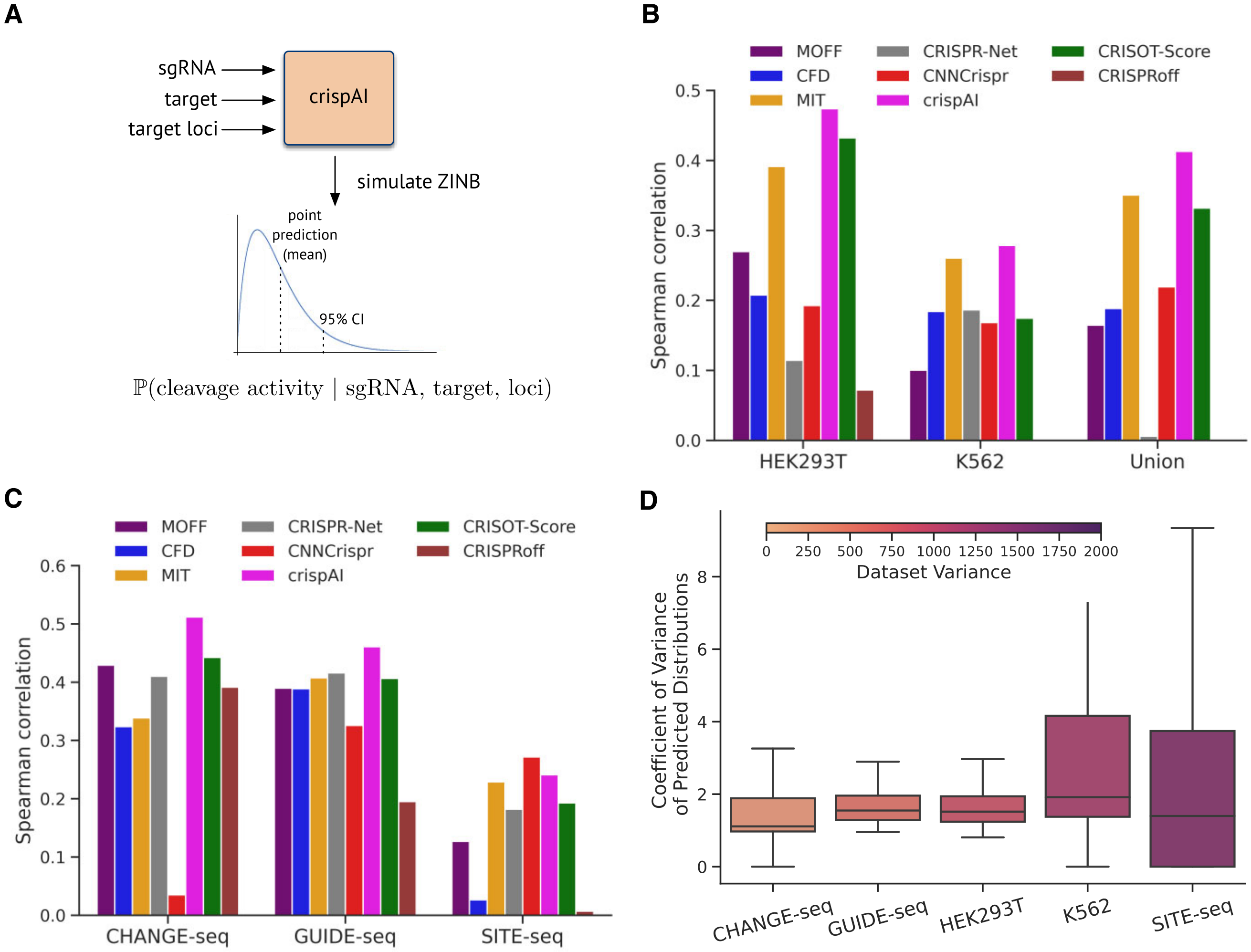


Figure 4. Off-target cleavage effect prediction with crispAI. **(A)** Input/Output schematic representation of crispAI. 23 −bp sgRNA and target DNA sequences with target DNA coordinates of the interface are inputs to our model. The model is trained to predict ZINB associated posterior parameters conditioned on the inputs. The conditional posterior distribution is then sampled. **(B)** Predictive performance comparison of competing models on HEK293T, K562 cell-line datasets and their unions comprising of $n = 536$, 120 and 656 samples respectively. Bar-plot displaying Spearman correlation coefficient between the ground truth cleavage values in the respective datasets and the predictions of MOFF, CFD, MIT, CRISPR-Net, CNNCrispr, CRISOT-Score, CRISPRoff scores and crispAI score. For performance comparison, expected value of crispAI-predicted posterior distributions are used as point predictions. **(C)** Similarly to part (B) performance of aforementioned methods are compared on CHANGE-seq, GUIDE-seq and SITE-seq datasets comprising of $n = 168$, 465, 443 and 6097 samples respectively. **(D)** Box-plots represent coefficient of variation for each predicted posterior distribution in all test datasets. Colorbar represents variances of CHANGE-seq, GUIDE-seq, HEK293T, K562 and SITE-seq datasets.

We observed a similar trend for the ground truth CHANGE-seq detected activity values for four different coefficient of variation intervals (i.e., [0,10], [10,20], [20,30], [30,40]), where the detected frequency values for crispAI-predicted distributions with lower coefficient of variation values are higher than those with higher coefficient of variation values. (Figure 5B) The relationship between ground truth CHANGE-seq detected off-target cleavage activity and the coefficient of variation of the crispAI-predicted distributions implies that as off-target cleavage activity decreases, our model's confidence in its predictions for the associated sgRNA-target pair increases. This phenomenon is advantageous as it indicates that crispAI is more certain and consistent in identifying sgRNA-target pairs with lower off-target cleavage activity. Such sgRNA–target pairs exhibit more densely distributed predicted probability distributions, suggesting a greater level of confidence in the predictions.

To investigate the effect of location and type of the mismatches on the uncertainty of the off-target cleavage activity, we stratified crispAI-predicted distributions on the test portion of the CHANGE-seq dataset with respect to: (i) the type of the mismatch (e.g., A→C—sgRNA base is A and target base is C); (ii) and the position of the mismatch between sgRNA and target DNA sequences. Then we plotted the coefficient of variation of the crispAI-predicted distributions for the stratified folds in Figure 5. We observed an increase in coefficient of variation for PAM-proximal base loci as opposed to PAM-distal region. Additionally, we observed a significant increase in uncertainty for some mismatch types compared to others. Specifically: T → G, T → C and C → G sgRNA to target mismatch types. Recent studies widely reported that PAM-proximal mismatches are less tolerated for the cleavage activity (6,42,43) meaning that PAM-proximal sequence-based variations have higher effect (Supplementary Table S3).

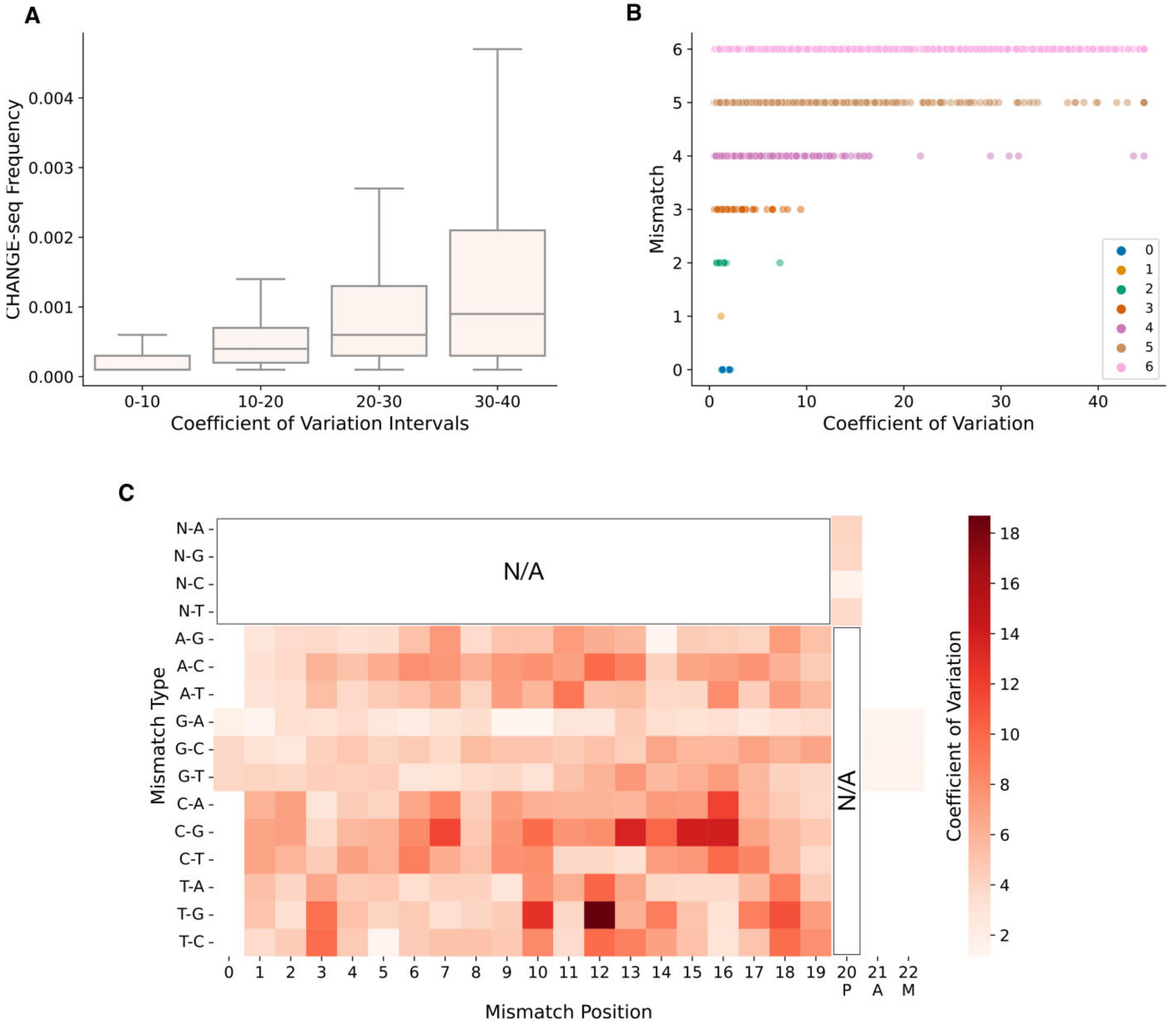


Figure 5. Similar to Figure 4D coefficient of variation for each crispAI-predicted posterior off-target activity distribution on the test portion of the CHANGE-seq dataset are plotted against ground truth CHANGE-seq detected off-target activity and number of mismatched positions between sgRNA and DNA sequences of the associated sample. **(A)** CHANGE-seq detected off-target activity values are depicted with box-plots for associated coefficient of variation intervals. Box-plots show the increase in CHANGE-seq detected off-target activity is associated with increase in coefficient of variation. **(B)** Scatter plot illustrating the relationship between coefficient variation and number of mismatched positions between sgRNA–target pair indicating higher uncertainty as the number of mismatches increase. **(C)** Grid-plot shows the coefficient of variation of the crispAI-predicted off-target cleavage activity distributions stratified with respect to types and positions of mismatches between sgRNA-target pairs on the test portion of the CHANGE-seq dataset. Certain mismatch types, such as T→G, and PAM-proximal mismatches yielded higher uncertainty.

PAM-proximal mismatches yielded higher uncertainty in the crispAI-predicted cleavage activity distributions.

crispAI-aggregate score enables uncertainty aware genome-wide sgRNA specificity prediction

We developed crispAI-aggregate, the first-of-its-kind genome-wide sgRNA specificity score, to provide aggregate score distributions for the sgRNA of interest. To calculate crispAI-aggregate distributions, we use Cas-OFFinder (27) to search for potential off-target sites of the sgRNA of interest up-to *N* mismatches, where *N* is a hyper-parameter of the crispAI-aggregate score. Then the element-wise summation of crispAI-predicted posterior cleavage activity distributions for all obtained off-target sites are element-wise divided by the crispAI-predicted posterior cleavage activity distribution for the perfect homology target-site sequence (i.e. 0-mismatch target). Finally, crispAI-aggregate is defined as the logarithm of the obtained conditional distribution (Materials and methods).

To evaluate crispAI-aggregate score, firstly we obtained the sgRNA specificity data curated by Fu *et al.* (19) on CHANGE-seq, TTISS (44) and GUIDE-seq datasets - providing specificity scores for 108, 59 and 10 sgRNAs respectively. We used sgRNAs in these dataset as inputs to crispAI-aggregate pipeline with maximum number of mismatches up-to $N = 5$ and obtained crispAI-aggregate distributions for all sgRNAs. To compare the performance of crispAI-aggregate score with other competing aggregation methods, we take expectations of the predicted distributions and obtained point predictions for the specificity aggregation task. Bar-plots in Figure 6A illustrates the performance of competing aggregate scores: MOFF-aggregate, CRISPR-Net, CFD, Elevation-aggregate, CRISPRspec, CRISOT-spec, CNN_std and crispAI-aggregate against Spearman correlation coefficient with respect to ground truth specificity values given in the curated dataset calculated based-on ground truth sequencing reads from respective *in vitro* assays. We observed that crispAI-aggregate significantly out-performed existing scores on CHANGE-seq and GUIDE-seq datasets with 14.27% and 6.99% improvements over best performing tools, MOFF and CRISPR-Net respectively and performed above average with 25.39% deterioration below MOFF-aggregate in Spearman correlation on sgRNAs in TTISS dataset. The average deterioriation among other methods is 41.26% below MOFF-aggregate score in Spearman correlation on this dataset (Supplementary Table S4).

Additionally, we visualised the Cumulative Distribution Functions (CDF) of the obtained crispAI-aggregate distributions for all 108 sgRNAs in the CHANGE-seq dataset in Figure 6B and annotated the CDFs with a colorbar depending on the associated ground truth specificity value. We observed that sgRNAs exhibiting higher specificity yielded crispAI-aggregate distributions with CDFs where the distribution is more densely populated around lower crispAI-aggregate score values (right-skewed PMFs). Whereas for the sgRNAs exhibiting lower specificity yielded crispAI-aggregate distributions with CDFs where the distribution is more densely populated around higher crispAI-aggregate score values (left-skewed PMFs). This result supports our findings since lower crispAI-aggregate score values indicate higher specificity in sgRNAs.

To further evaluate the crispAI-aggregate score, we obtained all 2408 sgRNAs presented in the Avana library (12) targeting non-essential genes. Similarly to Figure 6, we obtained crispAI-aggregate distributions for all sgRNAs, again using $N = 5$. Then, we created 8 bins in total using the ground-truth Log Fold Change (LFC) values present in the Avana library obtaining bins from −2.3 to 0.9 LFC values associated with all analysed sgRNAs. The ridgeline plot in Figure 7A illustrates the expected values of obtained crispAI-aggregate distributions for each bin. Similarly to Figure 6B for sgRNAs with higher LFC values expected values of crispAI-aggregate distributions are more skewed and more densely populated around lower crispAI-aggregate score values. More specifically for the obtained bins we observed 6.092, 5.166, 4.454, 4.069, 3.373, 2.821, 2.840 and 3.333 mean crispAI-aggregate scores with −2.14, −1.66, −1.28, −0.89, −0.48, −0.09, 0.19 and 0.5 average LFC values respectively. Additionally, we measured MOFF-aggregate score for the same bins, bar plot in Figure 7B shows the average MOFF-aggregate scores 4.148, 3.575, 2.880, 2.266, 1.203, 0.234, 0.076 and 0.958 for each LFC bin in the respective order (Supplementary Table S5).

The genome-wide specificity distributions produced by crispAI-aggregate enables distinguishing between sgRNAs with similar point predictions. To demonstrate this histograms in Figure 7C depicts crispAI-aggregate score distributions for two sgRNAs targeting CSH1 and SPACA7 genes with LFC values −0.305 and −0.294 in the Avana library and the mean values of crispAI-aggregate scores are calculated as 4.055 and 4.012, respectively. Although both the associated LFC values and the expected values of the predicted crispAI-aggregate distributions are very close for these sgRNAs, obtained distributions are different. Specifically, the sgRNA targeting CSH1 gene yielded a wider crispAI-aggregate score distribution compared to the sgRNA targeting SPACA7 gene with coefficient of variation values of 0.804 and 0.334. This finding suggests crispAI-aggregate score enables prioritization among sgRNAs with similar point predictions by providing richer information for genome-wide specificity prediction problem.

To set a brief example of the usefulness of the crispAI-aggregate score in clinical settings, we applied it to an experimental validation example reported by (37). This study focused on two therapeutically relevant genes, PCSK9 and BCL11A, which are targets for cholesterol reduction and treatment of β-thalassemia and sickle cell disease, respectively. We obtained crispAI-aggregate score distributions for both wild-type and modified versions of sgRNAs targeting these genes. For the PCSK9 target, we analyzed the wild-type sequence CGTGCGCAGGAGGACGAGGACGG and its modified version CGTGCGCAGTAGGACGAGGACGG (G10→T). For the BCL11A target, we examined the wild-type sequence GGCGAGACATGGTGGGCTGCGGG and its modified version GGCGAGACATTGTGGGCTGCGGG (G11→T). Our analysis, presented in Supplementary Figure S3, shows that the crispAI-aggregate scores for the modified sgRNAs improved significantly compared to their wild-type counterparts. Specifically, we observed a 193% improvement for the PCSK9 target and an 85% improvement for the BCL11A target. These results align well with the CRISOT-Spec scores reported by Chen *et al.*, which showed improvements of 108% and 92% for PCSK9 and BCL11A, respectively. Notably, Supplementary Figure S3 illustrates that the crispAI-aggregate generated posterior genome-wide specificity score distributions are right-skewed for the optimized (mutated) sgRNA sequences compared to the wild-type versions. This rightward skew indicates higher specificity for the optimized sgRNA sequences, providing a visual representation of the improvement in off-target effects. These findings demonstrate the capability of crispAI-aggregate to quantify improvements in sgRNA specificity, aligning with experimental validations. The ability to generate distributions rather than single-point estimates provides richer information for assessing genome-wide specificity, enabling more nuanced comparisons between different sgRNA designs in clinical settings as well.

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 (12,26,39,45), 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 (14,46–48). 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

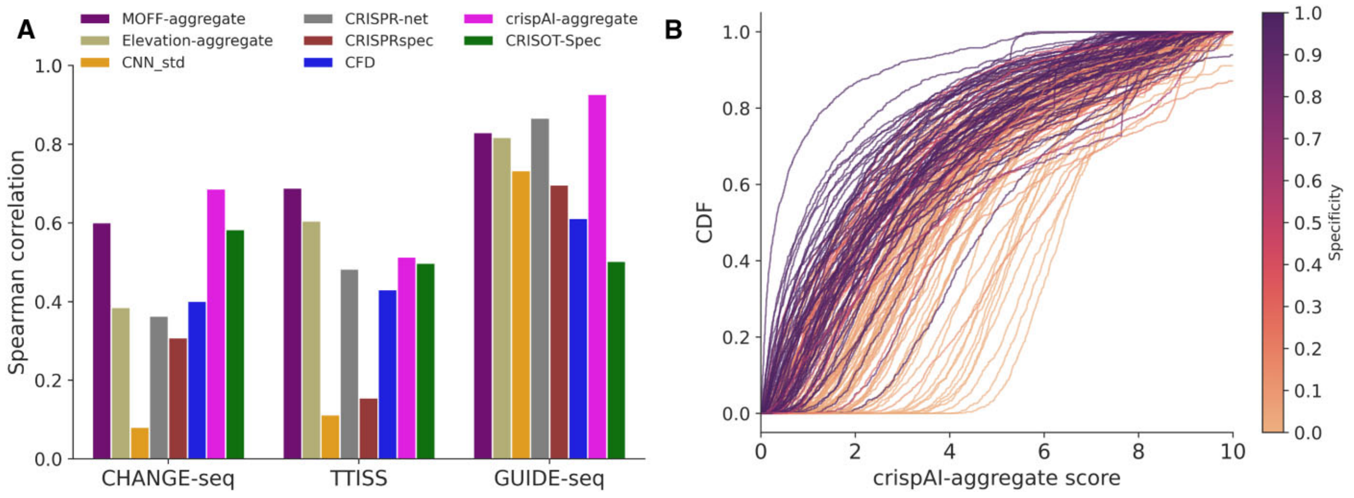


Figure 6. Genome-wide uncertainty aware sgRNA specificity prediction with crispAI-aggregate score. **(A)** Genome-wide sgRNA specificity score, crispAI-aggregate score, is defined as the logarithm of the ratio between sum of crispAI-predicted off-target scores up-to *N* mismatches and the on-target sequence, where *N* is a hyperparameter of the score. For the plots herein *N* = 5 is used. (A) Bar-plot represents the Spearman correlation between sgRNA specificity, as presented in (19), and predicted aggregate scores by MOFF-aggregate, CRISPR-Net, CFD, Elevation-aggregate, CRISOT-spec, CRISPRspec, CNN_std and crispAI-aggregate on CHANGE-seq, TTISS and GUIDE-seq datasets with *n* = 108, 59 and 10 sgRNAs respectively. **(B)** Cumulative Distribution Functions (CDFs) of predicted crispAI-aggregate score distributions are depicted. The colorbar represents the sgRNA specificty of the associated CDF in the respective dataset.

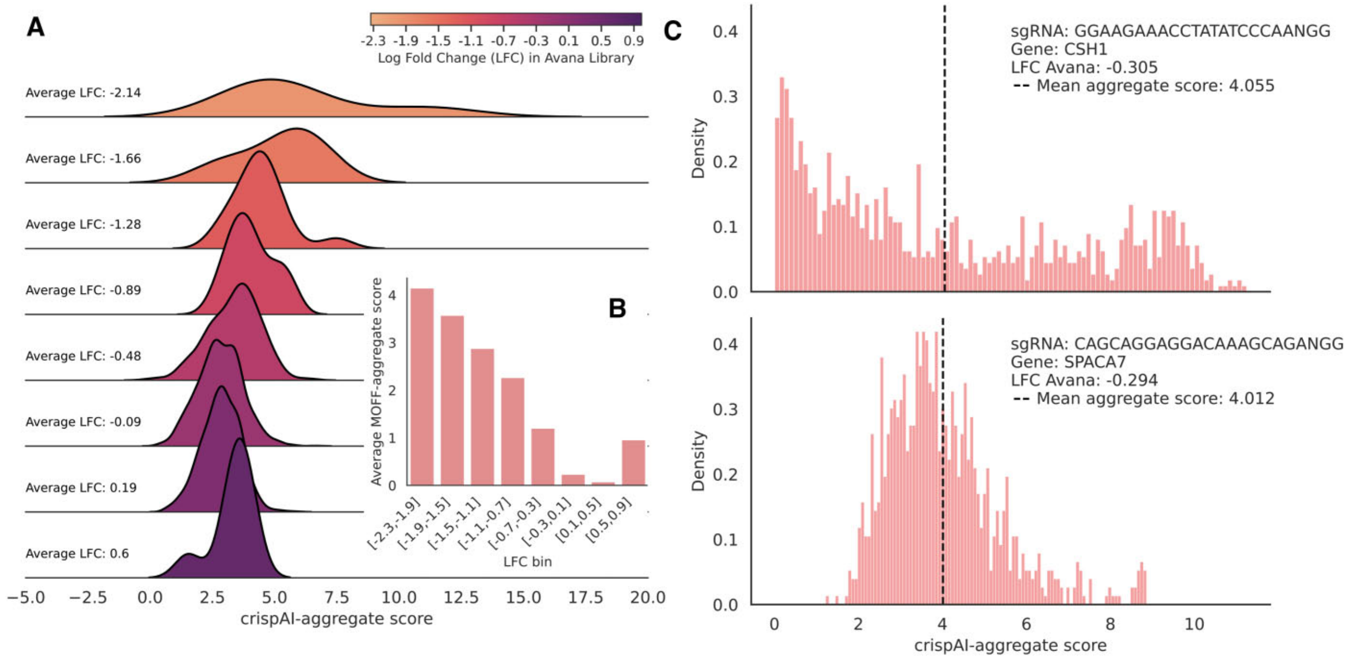


Figure 7. Prioritization of sgRNA specificity with crispAI-aggregate score. **(A)** Ridgeline plot for expected values of crispAI-aggregate score distributions on 2408 sgRNAs given in the Avana library (12). First, we searched genome-wide for up to 5 mismatch off-target sites for all 2408 sgRNAs using CasOffFinder, obtaining *n* samples in total, and calculated crispAI-aggregate scores for each sgRNA. Then, we used associated Log Fold Change (LFC) values for each sgRNA to obtain 8 bins with different bin ranges based on LFC values. The ridgeline plot depicts expected crispAI-aggregate score for each LFC bin. **(B)** Bar-plot representing the average MOFF-aggregate scores for the bins used in (A). **(C)** Histograms depict crispAI-aggregate score for two similar LFC sgRNAs targeting CSH1 and SPACA7 genes, with −0.305 and −0.294 LFC values in the Avana Library respectively. The expected value of predicted crispAI-aggregate score distributions are 4.055 and 4.012 in the respective order.

capabilities of the modelling efforts significantly improved (16,29,30,49). 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 (17). 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.

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 (23). 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 (50) studies, single-cell RNA sequencing (24,51) and bulk RNA sequencing (52), enabling improved accuracy and insight in differential gene expression analysis (53) 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 count noise-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 the abundant number of potential off-target sites, which is not suitable for modelling uncertainty in the on-target cleavage activity. Hence, a limitation of crispAI is that it is not designed for on-target activity prediction. The current model does not share 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. Whereas, crispAI incorporates physical features of the genomic loci at the target region, using richer information compared to sequence-based models and allowing differentiation between target regions with the same off-target sequence. Therefore, our method differentiates from the other uncertainty aware Gaussian Process Regression (GPR)-based method (22) by sampling uncertainty estimates using physical descriptors in addition to sequence-based features of the sgRNA-target context whereas the GPR-based model is unable to distinguish between target sites with the same off-target sequence.

We observed an increase in the coefficient of variations for the predicted activity distribution of sgRNA—target DNA interfaces that had mismatches in the PAM-proximal region compared to the coefficient of variation we observed for predicted activity distribution for interfaces with PAM-distal mismatches. This observed difference in the coefficient of variation for the predicted distributions for the said interfaces is in concordance with the different roles of PAM-proximal and PAM-distal base pairings have in the mechanism of the CRISPR/Cas9 based editing. Correct PAM-proximal base pairing is essential for initiating sgRNA-target DNA heteroduplex formation and therefore the stable binding of the CRISPR/Cas9 complex to the target DNA loci. Therefore PAM-proximal mismatches can adversely affect correct binding and subsequently off-target activity at the given loci (54). At the same time initiation of PAM-proximal base pairing is a stochastic process and this may explain why it is more challenging to correctly predict the effect of PAM-proximal mismatches with high certainty (55).

Data availability

All necessary scripts and data to replicate the results presented in figures are deposited to Zenodo https://doi.org/10.5281/zenodo.12609337, the tool and the trained model are available at https://github.com/furkanozdenn/crispr-offtarget-uncertainty and https://doi.org/10.5281/zenodo.13335960.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

In graphical abstract icons made by orvipixel, Freepik and pojok d from flaticon.com.

Author contributions: F.O. and P.M. designed the study. F.O. designed and implemented the model and performed the experiments. F.O. and P.M. wrote the manuscript.

Funding

Google DeepMind. Funding for open access charge: Google DeepMind.

Conflict of interest statement

None declared.

Author information

Peter Minary is a Research Lecturer at the Department of Computer Science, University of Oxford. His research interests include computational (structural) biology, machine learning and CRISPR-based genome editing technologies.

Furkan Ozden is a DPhil student in Computer Science at the University of Oxford. His research interests include machine learning, genomic variation and CRISPR-based genome editing technologies.

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.

References

1. Bhaya,D., Davison,M. and Barrangou,R. (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.*, **45**, 273–297.
2. Hsu,P.D., Lander,E.S. and Zhang,F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.
3. 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**, 84–87.
4. Barrangou,R. and Doudna,J.A. (2016) Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.*, **34**, 933–941.
5. 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 Res.*, **24**, 132–141.
6. 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. *Mol. Ther.-Nucleic Acids*, **4**, e264.
7. Mak,J.K., Störtz,F. and Minary,P. (2022) Comprehensive computational analysis of epigenetic descriptors affecting CRISPR-Cas9 off-target activity. *BMC Genom.*, **23**, 805.
8. 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. *Nat. Biotechnol.*, **33**, 187–197.
9. 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. *Nat. Biotechnol.*, **38**, 1317–1327.
10. 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. *Nat. Methods*, **12**, 237–243.
11. 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. *Nat. Methods*, **14**, 607–614.
12. 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. *Nat. Biotechnol.*, **34**, 184–191.
13. Chen,S.-A.A. and Tran,E. (2019) Optimizing precision genome editing through machine learning. *Forest*, **85**, 1–39.
14. 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. *Nat. Biomed. Eng.*, **2**, 38–47.
15. 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**, 1108–1115.
16. 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**, 51.
17. Störtz,F., Mak,J. and Minary,P. (2023) piCRISPR: physically informed deep learning models for CRISPR/Cas9 off-target cleavage prediction. *Artif. Int. Life Sci.*, **3**, 100075.
18. 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 Computat. Biol.*, **15**, e1007480.
19. Fu,R., He,W., Dou,J., Villarreal,O.D., Bedford,E., Wang,H., Hou,C., Zhang,L., Wang,Y., Ma,D., *et al.* (2022) Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity. *Nat. Commun.*, **13**, 474.
20. 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. *Brief. Bioinform.*, **24**, bbad131.
21. Gao,Y., Chuai,G., Yu,W., Qu,S. and Liu,Q. (2020) Data imbalance in CRISPR off-target prediction. *Brief. Bioinform.*, **21**, 1448–1454.
22. Kirillov,B., Savitskaya,E., Panov,M., Ogurtsov,A.Y., Shabalina,S.A., Koonin,E.V. and Severinov,K.V. (2022) Uncertainty-aware and interpretable evaluation of cas9–grna and cas12a–grna specificity for fully matched and partially mismatched targets with deep kernel learning. *Nucleic Acids Res.*, **50**, e11.
23. Silverman,J.D., Roche,K., Mukherjee,S. and David,L.A. (2020) Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J*, **18**, 2789–2798.
24. 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. *Nat. Commun.*, **10**, 390.
25. 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**, 1608.
26. 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 Res.*, **42**, W401–W407.
27. 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**, 1473–1475.
28. Concordet,J.-P. and Haeussler,M. (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.*, **46**, W242–W245.
29. 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 biol.*, **19**, 80.
30. 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. *Adv. sci.*, **7**, 1903562.
31. Zenil,H. and Minary,P. (2019) Training-free measures based on algorithmic probability identify high nucleosome occupancy in DNA sequences. *Nucleic Acids Res.*, **47**, e129.
32. Xi,L., Fondufe-Mittendorf,Y., Xia,L., Flatow,J., Widom,J. and Wang,J.-P. (2010) Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics*, **11**, 346.
33. Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. arXiv doi: https://arxiv.org/abs/1412.6980, 22 December 2014, preprint: not peer reviewed.
34. 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. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. pp. 8026–8037.
35. 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. *Nat. Biotechnol.*, **40**, 163–166.
36. Yan,J., Xue,D., Chuai,G., Gao,Y., Zhang,G. and Liu,Q. (2020) Benchmarking and integrating genome-wide CRISPR off-target detection and prediction. *Nucleic Acids Res.*, **48**, 11370–11379.
37. Chen,Q., Chuai,G., Zhang,H., Tang,J., Duan,L., Guan,H., Li,W., Li,W., Wen,J., Zuo,E., *et al.* (2023) Genome-wide CRISPR off-target prediction and optimization using RNA-DNA interaction fingerprints. *Nat. Commun.*, **14**, 7521.

38. Alkan,F., Wenzel,A., Anthon,C., Havgaard,J.H. and Gorodkin,J. (2018) CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome biol.*, **19**, 177.
39. 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. *Nat. Biotechnol.*, **31**, 827–832.
40. Kuleshov,V., Fenner,N. and Ermon,S. (2018) Accurate uncertainties for deep learning using calibrated regression. In: *International conference on machine learning*. PMLR, pp. 2796–2804.
41. Cameron,P., Fuller,C.K., Donohoue,P.D., Jones,B.N., Thompson,M.S., Carter,M.M., Gradia,S., Vidal,B., Garner,E., Slorach,E.M., *et al.* (2017) Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nat. Methods*, **14**, 600–606.
42. Lin,Y., Cradick,T.J., Brown,M.T., Deshmukh,H., Ranjan,P., Sarode,N., Wile,B.M., Vertino,P.M., Stewart,F.J. and Bao,G. (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.*, **42**, 7473–7485.
43. Wu,X., Kriz,A.J. and Sharp,P.A. (2014) Target specificity of the CRISPR-Cas9 system. *Quant. Biol.*, **2**, 59–70.
44. Schmid-Burgk,J.L., Gao,L., Li,D., Gardner,Z., Strecker,J., Lash,B. and Zhang,F. (2020) Highly parallel profiling of Cas9 variant specificity. *Mol. Cell*, **78**, 794–800.
45. 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**, e0124633.
46. 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**, 80–84.
47. 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. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.
48. 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 Res.*, **25**, 1147–1157.
49. 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.
50. 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**, 27.
51. 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 biol.*, **17**, 75.
52. Zhu,L., Lei,J., Devlin,B. and Roeder,K. (2018) A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.*, **12**, 609.
53. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.
54. Hille,F. and Charpentier,E. (2016) CRISPR-Cas: biology, mechanisms and relevance. *Philos. T. R. Soc. B: Biol. Sci.*, **371**, 20150496.
55. Shvets,A.A. and Kolomeisky,A.B. (2017) Mechanism of genome interrogation: How CRISPR RNA-guided Cas9 proteins locate specific targets on DNA. *Biophys. J.*, **113**, 1416–1424.

Received: January 31, 2024. **Revised:** August 7, 2024. **Editorial Decision:** August 10, 2024. **Accepted:** August 23, 2024