communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-08275-6

A versatile CRISPR/Cas9 system off-target prediction tool using language model

Weian Du[1], Liang Zhao[2], Kaichuan Diao[3], Yangyang Zheng[4], Qianyong Yang[5], Zhenzhen Zhu[2], Xiangxing Zhu[1,6] & Dongsheng Tang[1,6]

Genome editing with the CRISPR/Cas9 system has revolutionized life and medical sciences, particularly in treating monogenic genetic diseases by enabling long-term therapeutic effects from a single intervention. However, the CRISPR/Cas9 system can tolerate mismatches and DNA/RNA bulges at target sites, leading to unintended off-target effects that pose challenges for gene-editing therapy development. Existing high-throughput detection and in silico prediction methods are often limited to specifically designed single guide RNAs (sgRNAs) and perform poorly on unseen sequences. To address these limitations, we introduce CCLMoff, a deep learning framework for off-target prediction that incorporates a pretrained RNA language model from RNAcentral. CCLMoff captures mutual sequence information between sgRNAs and target sites and is trained on a comprehensive, updated dataset. This approach enables accurate off-target identification and strong generalization across diverse NGS-based detection datasets. Model interpretation reveals the biological importance of the seed region, underscoring CCLMoff's analytical capabilities. The development of CCLMoff lays the foundation for a comprehensive, end-to-end sgRNA design platform, enhancing both the precision and efficiency of CRISPR/Cas9-based therapeutics. CCLMoff is a versatile tool and is publicly available at github.com/duwa2/CCLMoff.

The CRISPR/Cas9 systems have been used to investigate target genes in genome modification[1], transcription[2] and splicing[3], and have been applied in various research settings to investigate and treat multiple genetic diseases[4,5], infectious diseases[6], immunological diseases[7], and cancers[8]. Among the exciting advances, the translational use of CRISPR/Cas9 system in monogenic human genetic diseases has the potential to provide long-term therapy after a single treatment[9]. However, extensive studies have also demonstrated that multiple mismatches as well as DNA/RNA bulges can be tolerated, resulting in the cleavage of unintended genomic sites, termed off-targets[10]. The potential off-target effect of the CRISPR/Cas9 system can lead to inadvertent gene-editing outcomes and become a bottleneck in the development of gene therapy[9].

Generally, the off-target effect is hard to discover as a result of the extremely low editing rate. Several experimental approaches have been developed to detect the off-target activity of CRISPR/Cas9 system. To provide clarity, the experimental detection techniques are divided into three major categories: (i) detection of Cas9 binding, such as Extru-seq[11], SELEX and its derivatives[12]; (ii) detection of Cas9-induced Double Strand Breaks (DSBs), such as in vitro techniques, Digenome-seq[13], CIRCLE-seq[14] and in vivo approaches DISCOVER-seq[15]; (iii) detection of repair products arising from Cas9-induced DSBs including IDLV[16] and GUIDE-seq[17]. However, while various experimental detection approaches can validate the defined sgRNA off-target effects, they fail to provide prior knowledge for sgRNA design. Computational methods address this limitation by utilizing the comprehensive datasets generated by these NGS-based approaches to construct predictive models, which efficiently forecast sgRNA off-target effects and offer valuable guidance for sgRNA design[18].

Recently, a variety of in silico tools for off-target prediction have been proposed. Based on their underlying principles, these methods can be categorized into four major groups[19]: (i) The alignment-based approach was

[1]Gene Editing Technology Center of Guangdong Province, School of Medicine, Foshan University, Foshan, Guangdong, China. [2]Shenzhen Health Development Research and Data Management Center, Shenzhen, Guangdong, China. [3]Shenzhen Center for Chronic Disease Control, Shenzhen, Guangdong, China. [4]Guangdong Homy Genetics Ltd, Foshan, Guangdong, China. [5]Jiujiang Key Laboratory of Rare Disease Research, Jiujiang University, Jiujiang, Jiangxi, China. [6]Guangdong Provincial Key Laboratory of Animal Molecular Design and Precise Breeding, Foshan University, Foshan, Guangdong, China.

e-mail: zhaoj93@mail2.sysu.edu.cn; zhu_xiangxing@126.com; tangdsh@163.com

the first computational method to introduce mismatch pattern into off-target prediction, such as Cas-OFFinder[20], CHOPCHOP[21] and GT-Scan[22]. These approaches employed different alignment methods to improve genome-wide scanning efficiency. (ii) Formula-based methods such as CCTop[23] and MIT[24] assigned the different mismatch weights of PAM-distal region and PAM-proximal region to aggregate the contribution of mismatch in different positions. (iii) Energy-based methods, including CRISPRoff[25], present an approximate binding energy model for the Cas9-gRNA-DNA chimeric complex. (iv) Learning-based methods, such as DeepCRISPR[26] and CRISPR-Net[27], can automatically extract the sequence information from training dataset to determine the genomic pattern of the off-target site. The deep learning-based methods exhibit superior performance and now serve as the state-of-the-art model in the off-target effect prediction[28]. However, existing deep learning-based models are often trained on limited datasets containing a small number of sgRNAs and NGS-based off-target detection data, which restricts their generalization ability and confines their applicability to specific detection approaches.

To address these limitations, we proposed a deep learning framework, namely CCLMoff, which incorporates the RNA language model to extract the sequence information and the genomic contexts. Besides, we compiled a comprehensive dataset with 13 genome-wide off-target detection technologies, forcing CCLMoff to learn the general off-target pattern. Thus, CCLMoff demonstrated superior performance over the state-of-the-art model in various scenarios. The thorough evaluation showed that CCLMoff accurately identified off-target sites and displayed strong cross-dataset generalization ability. The model interpretation analysis indicated that CCLMoff successfully captured the seed region for off-target prediction. The development of CCLMoff paves the way for the establishment of a versatile and end-to-end in silico sgRNA design platform.

Methods

Data source

In order to guide the construction of a universal and versatile model for off-target prediction, we first curated a comprehensive off-target dataset encompassing a wide range of validated sgRNAs and diverse off-target detection methods. We specifically excluded targeted site detection techniques such as targeted PCR, and focused on the genome-wide deep sequencing-based off-target detection approaches to ensure the model's capability to detect off-target sites on a genome-wide scale. In total, we integrated 13 genome-wide deep sequencing techniques from 21 publications, categorized into three groups based on their detection methods: DNA binding detection methods (Extru-seq[11], SITE-seq[29]); the DSB detection methods (CIRCLE-seq[14], DISCOVER-seq[15], DISCOVER-seq+[30], CHANGE-seq[31], BLESS[32]); the repair product detection methods (GUIDE-seq[17], Digenome-seq[13], DIG-seq[33], IDLV[34], HTGTS[35] and SURRO-seq[36]). However, these studies only released validated off-target sites corresponding to the tested sgRNAs. During the model training process, negative off-target sites need to be externally constructed.

To generate an appropriate negative dataset, Cas-OFFinder[20] was employed for the negative sample construction, imposing constraints on the number of mismatches and bulges to ensure a representative distribution between off-target sites and mismatch candidates. The negative dataset was divided into two major categories based on whether the corresponding positive off-target sites contained bulges. As only recent studies[27] account for bulge information, many earlier studies do not incorporate it. To ensure a fair comparison, we constructed two distinct negative datasets. For positive samples with bulge information, Cas-OFFinder was configured with parameters allowing up to 6 mismatches and 1 bulge. For positive samples without bulge information, Cas-OFFinder was set to consider only up to six mismatches between the sgRNA and the target sites. When constructing negative samples using Cas-OFFinder, CCLMoff was designed to identify off-target sites from mismatch candidates, effectively reducing the sampling space and providing challenging samples to enhance the model's ability to distinguish off-target sites.

Model construction

To address the off-target prediction problem, which has two components: sgRNA and target site, we adopted a question-answering framework, wherein we formulated the problem as follows (Fig. 1). The input to this framework consisted of two separate parts: the sgRNA sequence, which served as the question stem, and the target site candidate, which acted as the answer. The target site, being a DNA sequence, was transformed into pseudo-RNA by substituting thymine (T) with uracil (U) when using a language model pretrained on RNA. The primary objective was to ascertain

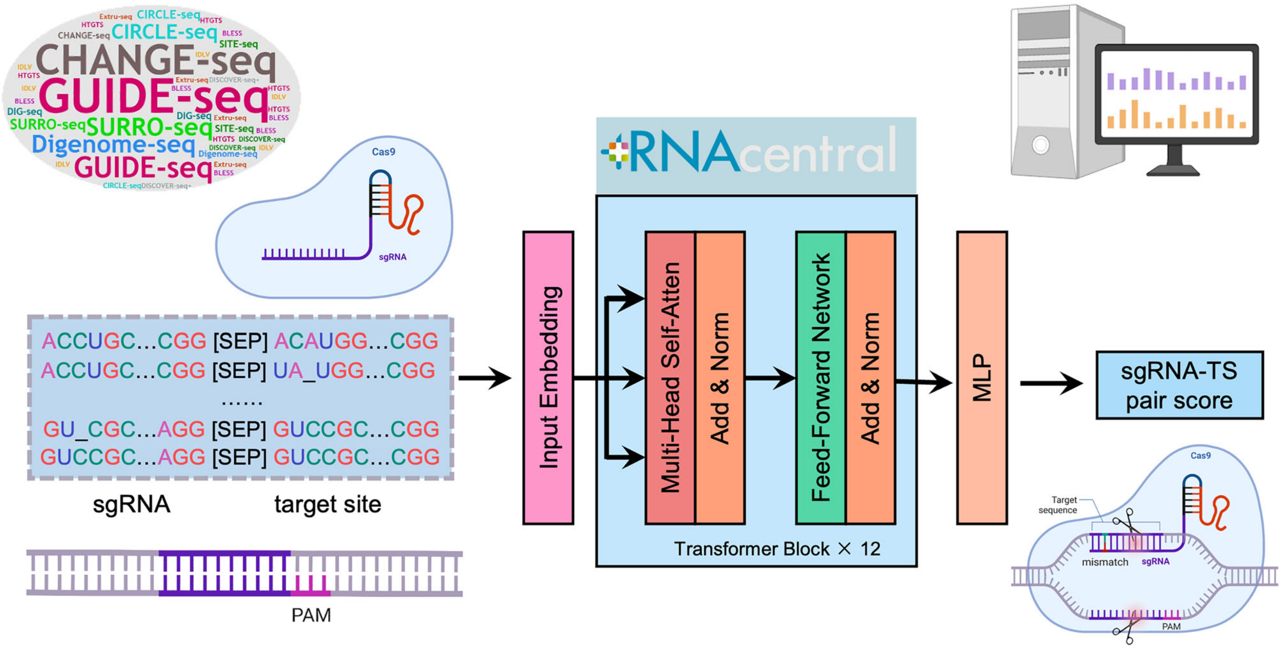

Fig. 1 | Overview of the pipeline for CCLMoff. The high-throughput off-target data is encoded as the sgRNA-target site pair and concatenated by a predefined token [SEP]. The sgRNA-target site pairs go through the Input Embedding layer and feed into 12 Transformer Blocks initialized by RNAcentral. The [CLS] of the final hidden layer is employed for classification using the multilayer perceptron to predict the sgRNA-target site pair score.

whether the sgRNA sequence could interact with the candidate pseudo-RNA sequence and result into off-target effect. Building upon this hypothesis, we developed a transformer-based language model to fulfill this classification task. Initially, we tokenized the sgRNA sequence and candidate pseudo-RNA sequence at the nucleotide level, using Input Embedding block. To indicate their discontinuity, we introduced a special token [SEP] as a delimiter to separate them. Subsequently, the input embeddings of the sgRNA and the pseudo-RNA candidate were input into an encoder, which is composed of 12 transformer blocks. These transformer blocks with a multi-head attention module enabled effective information processing and contextual feature extraction between sgRNA and the target site.

For the off-target classification task, we utilized the final hidden layer state of the transformer encoder. Specifically, the [CLS] token from the final hidden layer was employed as the input for a Multilayer Perceptron (MLP), which was tasked with predicting the sgRNA-target site pair score. This MLP layer generated a score representing the likelihood that the candidate pseudo-RNA sequence is an off-target site for the sgRNA. The scoring mechanism evaluated the compatibility and binding affinity between the sgRNA and the candidate pseudo-RNA sequence. The encoder, consisting of 12 transformer blocks, was initialized using the RNA-FM model[37], which had been pretrained on 23 million RNA sequences from RNAcentral[38]. This pre-training approach ensured that the encoder had a robust understanding of RNA sequences, enhancing its ability to capture relevant features during the subsequent off-target prediction task.

To evaluate the impact of epigenetic information on model performance, we incorporated epigenetic data obtained from DeepCRISPR. A convolutional neural network (CNN) was used to encode four epigenetic channels: CTCF binding information, H3K4me3 histone modification, chromatin accessibility, and DNA methylation derived from reduced representation bisulfite sequencing (RRBS). The resulting representation vector was then concatenated with the output of the language model and fed into the MLP layer. This enhanced model is referred to as CCLMoff-Epi. In addition, we introduced a model without the pretrained language model, referred to as CCLMoff-Vanilla, which was trained from scratch on the off-target dataset. This setup was designed to evaluate the impact of the pre-trained language model RNA-FM on model performance.

Training process. To utilize the robust feature extraction ability from the RNA-FM foundation model, we set a small learning rate for the parameter of 12 Transformer Blocks. The CCLMoff is trained using a standard binary cross-entropy (BCE) loss defined as follows:

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

where y_i represents the true label (off-target or not) for the i-th sample, and $\hat{y}_i$ denotes the predicted probability that the candidate pseudo-RNA is an off-target for the corresponding sgRNA. This loss function was used to guide the optimization of the model, ensuring that it effectively distinguished between off-target and non-off-target sequences. We employed the AdamW optimizer with a learning rate of 5×10^{-4} for the 12 Transformer encoder blockers parameters and 1×10^{-3} for the parameters of the MLP. A learning rate warm-up strategy was applied during the first 5 epochs. CCLMoff was trained for 10 epochs using 8 NVIDIA A100 80G GPUs, with a batch size of 128, and the total training time amounted to ~3 h. The deep learning baseline models, including CRISPR-Net, LSTM, CCLMoff-Epi and CCLMoff-Vanilla were trained using the consistent hyperparameters. Due to the dataset's high level of imbalance, the bootstrapping sampling strategy was applied to ensure an equal number of positive and negative samples in each training batch.

Interpretation analysis

Attention scores serve as a fundamental mechanism in attention-based models, particularly in language models, by dynamically determining the importance of different elements within an input sequence. This mechanism allows the model to focus selectively on relevant parts of the data. By assigning varying weights to different tokens, attention scores enable the model to capture contextual relationships, long-range dependencies, and nuanced interactions between sgRNA and target site. This selective weighting not only enhances the model's predictive capabilities but also provides a window into its decision-making process, making it a cornerstone for both performance and interoperability. Intuitively, these scores reflect the contribution of each nucleotide to the model's prediction, with higher attention scores indicating a greater significance of the nucleotide's position and composition in the sequence. The attention mechanism allows the model to focus on specific parts of the input, helping it prioritize key features relevant to off-target prediction. The attention score for each nucleotide is calculated as follows:

$$Attention\ Score = \frac{Q \times k_l}{\sqrt{d}}$$

where Q is the query vector, and k_l is the lth column of K. Note that the score functions presented in this section can be more efficiently calculated in matrix form using K instead of each column separately.

Statistics and reproducibility

All computational experiments were performed with clearly defined training and testing procedures. For model training, we used a benchmark dataset composed of 418 sgRNAs and 82,699 validated off-target sites, along with 9,521,638 negative samples generated via Cas-OFFinder. To address the data imbalance (positive-to-negative ratios ranging from 1:26 to 1:4189), we applied a bootstrapping sampling strategy that ensured an equal number of positive and negative samples in each training batch. Model performance was evaluated using standard metrics including balanced accuracy, F1-score, area under the Receiver Operating Characteristic curve (AUROC), and area under the precision–recall curve (AUPRC). Five-fold cross-validation was performed to assess model robustness, and all reported results represent the mean and standard deviation across five repeated runs with different random seeds. For statistical comparisons, two-sided Student's t tests were conducted to assess performance differences between CCLMoff and baseline models, with P values <0.05 considered significant. All comparisons were carried out using identical training settings, data partitions, and evaluation metrics to ensure fairness and reproducibility.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Benchmark dataset construction

In this study, we aimed to create a comprehensive off-target dataset to facilitate the development of a universal model for off-target prediction. To ensure the dataset's suitability for genome-wide off-target detection, we excluded targeted site detection methods, such as targeted PCR, and concentrated on genome-wide and deep sequencing-based off-target detection approaches. Thus, we incorporated 13 genome-wide deep sequencing techniques, including GUIDE-seq[17], CIRCLE-seq[14], SITE-seq[29], DISCOVER-seq[15], DISCOVER-seq+[30], CHANGE-seq[31], Digenome-seq[13], DIG-seq[33], HTGTS[35], IDLV[34], BLESS[32], Extru-seq[11], and SURRO-seq[36] to construct a comprehensive off-target dataset categorized into three groups based on the underlying sequencing mechanisms. A bunch of in silico model studies have curated datasets for off-target prediction, yet these datasets face significant limitations, such as covering a small number of sgRNAs or employing inconsistent criteria for constructing negative samples. For instance, the dataset utilized by CRISPR-Net[27] contained only 145 sgRNAs and relied on five deep sequencing-based off-target detection methods. For negative sample construction, CRISPR-Net employed Cas-OFFinder with a parameter setting of 6 mismatches and 1 bulge. Similarly, CrisprDNT[39] compiled datasets from multiple studies, comprising 149 sgRNAs, five deep sequencing approaches, and one PCR-based method, which is a targeted off-target detection technique. CrisprDNT also utilized Cas-OFFinder but with different parameters, permitting six mismatches without bulge information. This highlights the urgent need for researchers to establish a standardized benchmark dataset to unify and address the off-target site prediction challenge.

For the benchmark dataset, we collected the positive off-target sites from the 21 studies, including both the original publications and their application studies (Table 1). These studies only provided validated off-target sites for the tested sgRNAs, requiring the negative off-target sites to be externally constructed during the model training process. For the construction of the negative dataset, we categorized the data into two groups based on whether the positive off-target sites contained bulge information. For positive samples with bulge information, we set the Cas-OFFinder parameters to allow six mismatches and one bulge. For positive samples without bulge information, Cas-OFFinder was configured to consider up to six mismatches between the sgRNA and target sites.

In developing a universal model, our dataset includes two species and four reference genomes. Furthermore, uncanonical-length sgRNAs were incorporated to enhance the dataset's breadth and applicability. Overall, we constructed a benchmark dataset integrating 13 off-target detection technologies from 21 studies, making it the most comprehensive dataset to date, with 418 sgRNAs and 82,699 validated off-target sites. The negative dataset, generated using Cas-OFFinder with constrained parameters, consisted of a total of 9,521,638 negative samples. Given the highly imbalanced nature of the dataset, with imbalance ratios ranging from 26:1 to 4189:1, addressing this imbalance was a critical focus during the model construction and training process. In comparison to the recently published database CrisprSQL[40], which contains cleavage data from only 144 guide RNAs and 25,632 guide-target pairs, our dataset is significantly more comprehensive. Moreover, CrisprSQL only includes positive target sites and lacks negative sample information, making it insufficient for training predictive models. Furthermore, Sherkatghanad et al.[41] provide a comprehensive review of CRISPR/Cas-related computational challenges, including a table summarizing six off-target studies. This covers datasets such as GUIDE-seq and CHANGE-seq, which represent only a small fraction of the datasets incorporated into our benchmark. Consequently, our benchmark dataset offers a more complete resource for off-target prediction, addressing both positive and negative off-target sites to support future model development.

Language model improve the off-target prediction

To evaluate the capability of the pretrained language model framework, we adopted a rigorous training process identical to that of the baseline models. Both CCLMoff and the baseline model were trained on the same dataset to facilitate a direct performance comparison and to assess the effectiveness of CCLMoff framework relative to the baseline model. Thus, we conducted a thorough evaluation of CCLMoff on the CIRCLE-seq dataset and found that it exhibited superior performance in identifying off-target sites compared to SOTA models, including CRISPR-Net, LSTM and CCTop. Besides, we also incorporated two variants of CCLMoff called CCLMoff-Epi and CCLMoff-Vanilla, to investigate the impact of incorporating epigenetic information and pretraining process on massive RNA sequence datasets (Fig. 2). For a fair comparison, we only included the models capable of accounting for bulge information in this section. The results (Table 2) demonstrated the superior performance of the pretrained language model across various metrics, including balanced accuracy, F1-score, AUROC and AUPRC. CCLMoff outperformed the state-of-the-art models, achieving a balanced accuracy of 0.998 (± 0.001), an F1-score of 0.409 (± 0.003), an AUROC of

Table 1 | Summary of the comprehensive dataset construction with technique categorization based on the off-target detection mechanism

ID	Technique	Category	Total $Pos + Neg$	Target site	Imbal ratio $\frac{Pos+Neg}{Pos}$	sgRNA	Genome	Bulge	sgRNA length	Ref.
1	CIRCLE-seq	DSB detection	1,683,395	5796	290	11	hg19	Yes	19, 20	14
2	GUIDE-seq	Repair Product Detection	1,599,541	414	3864	10	hg19	Yes	20	17
3	GUIDE-seq	Repair Product Detection	195,186	414	471	10	hg19	No	20	17
4	DISCOVER-seq	DSB Detection	104,971	31	3386	4	hg19	No	20	15
5	DISCOVER-seq+	DSB Detection	105,038	98	1072	4	hg19	No	20	30
6	DISCOVER-seq	DSB Detection	26,757	49	546	1	mm10	No	20	15
7	DISCOVER-seq+	DSB Detection	26,805	98	274	1	mm10	No	20	30
8	CHANGE-seq	DSB detection	1,880,364	71,254	26	110	hg19	No	20	31
9	GUIDE-seq	Repair Product Detection	804,809	1702	473	58	hg19	No	20	31
10	Extru-seq	DNA Binding	205,375	94	2185	5	hg19	No	19	11
11	Extru-seq	DNA Binding	137,936	56	2463	2	mm10	No	19	11
12	SURRO-seq	Repair Product Detection	1,263,524	863	1464	110	hg19	No	20	36
13	SITE-seq	DNA Binding	102,691	89	1154	8	hg38	No	20	54
14	Digenome-seq	Repair Product Detection	25,796	162	159	2	hg19	Yes	20	13
15	Digenome-seq	Repair Product Detection	195,031	258	756	10	hg19	No	20	13
16	DIG-seq	Repair Product Detection	91,297	141	647	8	hg38	No	20	33
17	GUIDE-seq	Repair Product Detection	385,759	426	906	31	mm9	No	20, 21	58
18	GUIDE-seq	Repair Product Detection	84,703	61	1389	7	hg19	Yes	20	59
19	GUIDE-seq	Repair Product Detection	162,136	272	596	9	hg19	No	20	18
20	GUIDE-seq	Repair Product Detection	115,465	203	569	5	hg19	No	20	60
21	HTGTS	Repair Product Detection	49,178	87	565	3	hg19	No	20	35
22	IDLV	Repair Product Detection	54,459	13	4189	2	hg19	No	20	34
23	BLESS	DSB Detection	73,807	31	2381	2	hg19	No	20	61
24	BLESS	DSB Detection	113,291	53	2138	3	hg19	No	19, 20	61
25	BLESS	DSB Detection	34,324	34	1010	2	mm9	No	19, 20	62

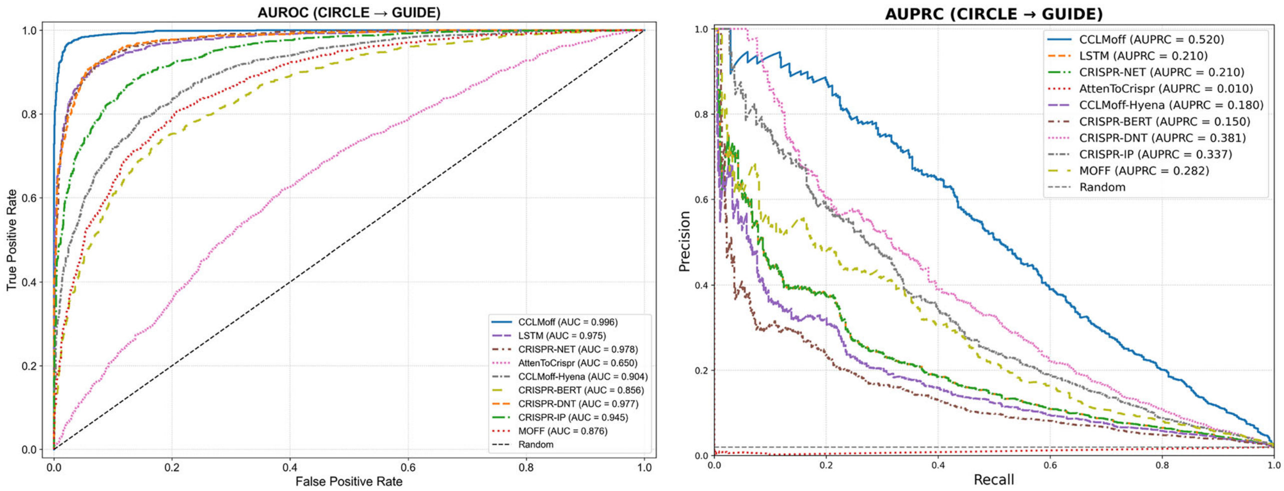

Fig. 2 | Cross-dataset evaluation on the GUIDE-seq dataset. The models were trained on the CIRCLE-seq dataset and externally validated on GUIDE-seq, which uses a different experimental platform. Left: ROC curves showing the AUROC performance of all models. Right**:** Precision–recall curves showing AUPRC performance, which is particularly important in imbalanced datasets. The RNA-pretrained model CCLMoff achieved the best performance (AUROC = 0.996, AUPRC = 0.520), significantly outperforming baseline models including LSTM, CRISPR-Net, and AttenToCrispr. In addition, we evaluated three recent language model-based approaches: CCLMoff-Hyena (DNA-pretrained), CRISPR-BERT (task-specific pretraining), and CRISPR-DNT (Transformer-based). Both CRISPR-BERT and CCLMoff-Hyena demonstrated lower AUPRC than CCLMoff, highlighting the advantage of RNA-specific foundation model pretraining in capturing sgRNA–DNA interactions.

Table 2 | Cross-validation results on the CIRCLE-seq dataset

Model	Bal Acc	F1-score	AUROC	AUPRC
CCLMoff	0.998 ± 0.001	0.409 ± 0.003	0.985 ± 0.001	0.524 ± 0.004
LSTM	0.843 ± 0.002	0.052 ± 0.001	0.926 ± 0.001	0.479 ± 0.003
CRISPR-Net	0.806 ± 0.002	0.083 ± 0.001	0.915 ± 0.003	0.462 ± 0.007
CCTop	0.887 ± 0.004	0.003 ± 0.001	0.711 ± 0.004	0.008 ± 0.001
CCLMoff-Epi	0.998 ± 0.001	0.429 ± 0.005	0.989 ± 0.001	0.513 ± 0.004
CCLMoff-Van	0.836 ± 0.001	0.053 ± 0.001	0.901 ± 0.003	0.422 ± 0.005
CCLMoff v.s.				
LSTM	6×10^{-12}	3×10^{-12}	9×10^{-9}	4×10^{-11}
CRISPR-Net	1×10^{-9}	8×10^{-7}	1×10^{-5}	7×10^{-6}
CCTop	1×10^{-8}	1×10^{-23}	1×10^{-13}	1×10^{-18}
CCLMoff-Epi	0.82	0.53	0.09	0.12
CCLMoff-Van	1×10^{-4}	2×10^{-13}	6×10^{-4}	5×10^{-6}

Bal Acc Balanced Accuracy, used for evaluating imbalanced datasets, *CCLMoff-Van* CCLMoff-Vanilla.
A *t* test was conducted to assess the statistical significance of performance differences between CCLMoff and baseline models.

0.985 (± 0.001), and an AUPRC of 0.524 (± 0.004). Notably, we observed that the simplified LSTM version of CRISPR-Net outperformed the original version (CRISPR-Net) in terms of balanced accuracy, AUROC, and AUPRC. This suggests that the RNN framework may be more effective in processing raw sequence data. The CCLMoff-Epi obtained a balanced accuracy of 0.998 (± 0.001), an F1-score of 0.429 (± 0.005), an AUROC of 0.989 (± 0.001), and an AUPRC of 0.513 (± 0.004). However, the addition of four channels of epigenetic information, including DNase, CTCF, H3K4me3, and RRBS, did not result in significant improvement. We hypothesize that the pretrained language model inherently captures epigenetic and genomic context information during training, rendering the extra epigenetic channels unnecessary for further enhancing off-target prediction performance. Notably, CCLMoff-Vanilla, trained from scratch without a pretraining process on a large-scale dataset, still achieved considerable performance, which we attribute to the strength of its transformer-based framework. The detailed AUROC and AUPRC figures for each sgRNA were shown in Supplementary Figs. S1 and S2. A *t* test further confirmed that CCLMoff significantly outperformed CRISPR-Net, LSTM, and CCTop across all metrics. These findings indicate that the language model framework effectively captures mutual information between the sgRNA and target site, resulting in superior performance in off-target prediction.

Language model exhibits robust cross-dataset generalization ability

To comprehensively evaluate the generalization ability of the pretrained language model in CCLMoff, we trained the model on a specific dataset and evaluated it on an external dataset from a different experimental category. Specifically, we trained CCLMoff on the CIRCLE-seq dataset and validated its performance on GUIDE-seq, which utilizes a distinct sequencing mechanism. The results demonstrated that CCLMoff exhibits robust cross-dataset generalization, significantly outperforming existing state-of-the-art models such as AttenToCrispr[42], CRISPR-Net[27], and LSTM. Notably, CCLMoff achieved an AUROC of 0.996 and an AUPRC of 0.520 on the GUIDE-seq dataset—substantially higher than the AUPRC of 0.210 attained by CRISPR-Net.

To further contextualize these findings, we compared CCLMoff with several recent and high-performing methods. The first is CRISPR-IP[43], a graph-based model that incorporates mismatch position encoding. Trained on CIRCLE-seq and tested on GUIDE-seq, CRISPR-IP achieved an AUROC of 0.945 and an AUPRC of 0.337. The second is MOFF[44], a random forest model that integrates chromatin accessibility and sequence features, which achieved an AUROC of 0.876 and an AUPRC of 0.282. The third is CRISPR-DNT[39], a dinucleotide-enhanced neural network model that obtained an AUROC of 0.977 and an AUPRC of 0.381 under the same evaluation setting. While these recent models demonstrate competitive performance, CCLMoff consistently outperforms them across evaluation metrics. We further benchmarked CCLMoff against two recent language model-based approaches to evaluate the impact of pretraining modality.

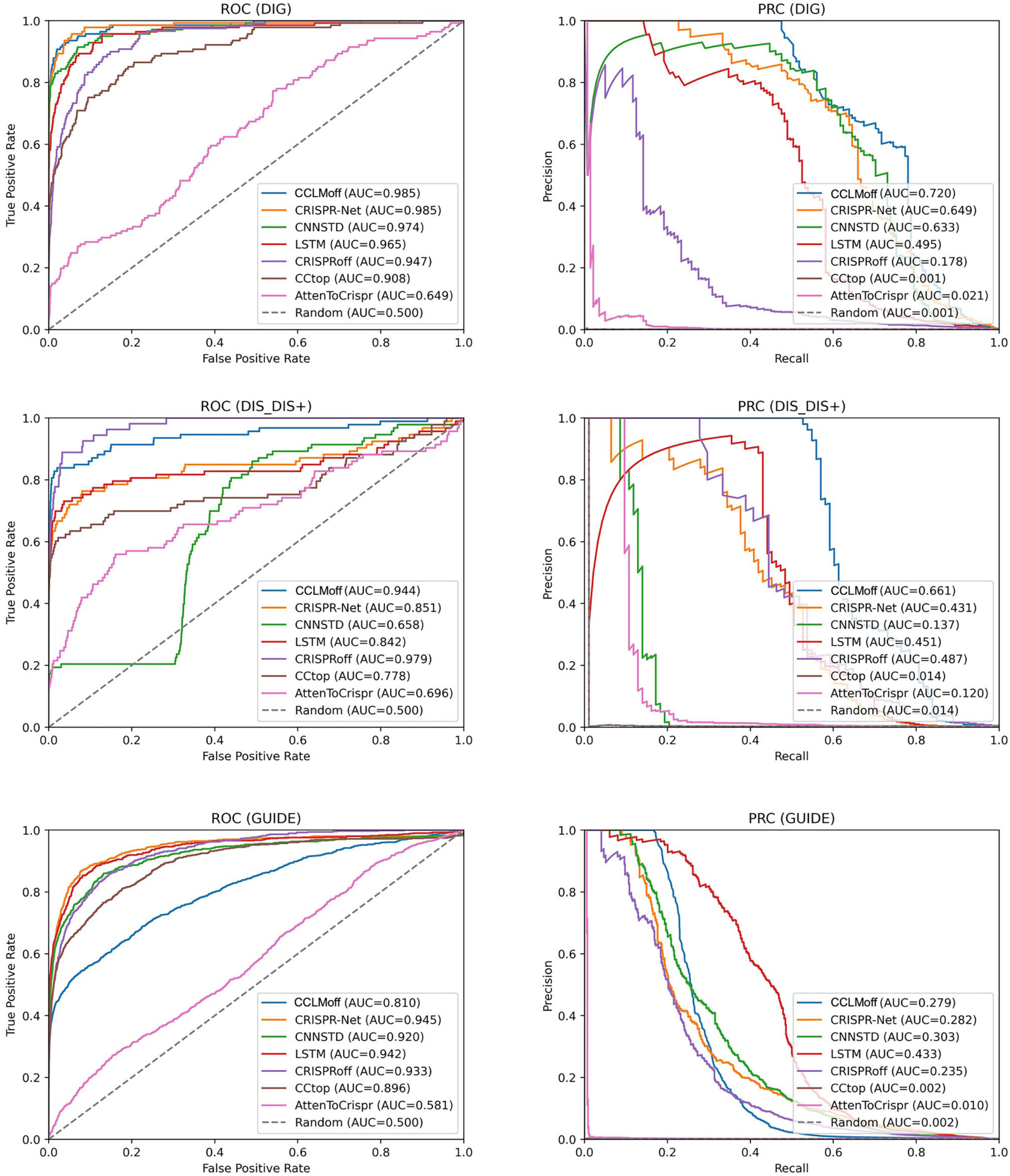

Fig. 3 | The model performance on external validation. On DIG-seq dataset, CCLMoff achieved superior performance (AUROC=0.985 and AUPRC=0.720) than the SOTA model, indicating that CCLMoff can successfully capture off-target pattern revealed by DIG-seq. In DISCOVER-seq and DISCOVER-seq+ dataset, CCLMoff exhibited superior performance in AUPRC (AUPRC=0.661) and considerable performance in AUROC (AUROC=0.944), indicating that CCLMoff have sufficient capacity in recalling the potential off-target sites. In GUIDE-seq dataset, CCLMoff exhibited limited performance (AUPRC=0.810, AUROC=0.279), due to the baseline model was directly trained on the dataset of GUIDE-seq, indicating that the existing model intend to be an approach-specific model instead of general off-target site prediction model.

The first is CCLMoff-Hyena, a variant of our model that incorporates the genomic DNA foundation model HyenaDNA[45]. This model was trained on CIRCLE-seq and tested on GUIDE-seq, achieving an AUROC of 0.904 and an AUPRC of 0.180. The second is CRISPR-BERT[46], a task-specific BERT model trained from scratch on sgRNA-target site pairs without large-scale DNA or RNA pretraining. Under the same protocol, CRISPR-BERT achieved an AUROC of 0.856 and an AUPRC of 0.150. These results underscore the effectiveness of RNA-specific pretraining for modeling

CRISPR-Cas9 off-target effects. Compared to CRISPR-BERT, which suffers from limited generalization due to its narrow training domain, and CCLMoff-Hyena, which is pretrained on genomic DNA rather than RNA, CCLMoff leverages large-scale RNA-based pretraining to better capture the biological properties of sgRNA and its interactions with DNA target sites. This is particularly important given that sgRNA is an RNA molecule, and modeling it in the correct modality appears to be beneficial for downstream prediction tasks. These findings demonstrate that CCLMoff captures shared off-target sequence patterns across datasets and technologies, enabling it to generalize across cell types, sequencing protocols, and experimental conditions. Furthermore, these results highlight the potential of pretrained language models in developing scalable and robust tools for genome editing, and lay a strong foundation for future extensions that integrate diverse data sources to further enhance off-target prediction.

CCLMoff can accurately predict off-target sites

The pretrained language model's robust generalization ability prompted us to train the full version of CCLMoff on a comprehensive dataset to learn the general patterns of off-target effect occurrence. We employed a leave-one-dataset-out evaluation strategy to assess the model's performance on the latest datasets: DIG-seq, GUIDE-seq, DISCOVER-seq, and DISCOVER-seq+. As these datasets are bulge-free, we included additional mismatch-only models, such as AttenToCrispr[42], CNN_std[47], and CRISPRoff[48] for performance comparison (Fig. 3). These models were evaluated using the same parameters as specified in their original papers and implementations available on GitHub (ref "Code availability").

CCLMoff achieved outstanding performance across all evaluated datasets. In the DIG-seq dataset, CCLMoff demonstrated superior results, with an AUROC of 0.985 and an AUPRC of 0.720, outperforming state-of-the-art models and showcasing its ability to capture off-target patterns revealed by DIG-seq. In the DISCOVER-seq and DISCOVER-seq+ datasets, CCLMoff also exhibited superior performance in terms of AUPRC (0.665) and solid performance in AUROC (0.944), demonstrating its effectiveness in recalling potential off-target sites. However, in the GUIDE-seq dataset, CCLMoff showed more considerable performance (AUPRC = 0.810, AUROC = 0.279), while the state-of-the-art models (e.g., CRISPR-Net and CNNSTD) used for comparison were directly trained on the GUIDE-seq dataset. CCLMoff used the leave-one-dataset-out strategy and was unable to assess GUIDE-seq dataset during the training process. This suggests that these baseline models may be more approach-specific rather than providing general off-target site prediction capabilities. Overall, these results highlight that CCLMoff trained on the comprehensive dataset can accurately predict off-target sites and capture general off-target occurrence patterns, making it a valuable tool for general off-target prediction across diverse datasets and technologies.

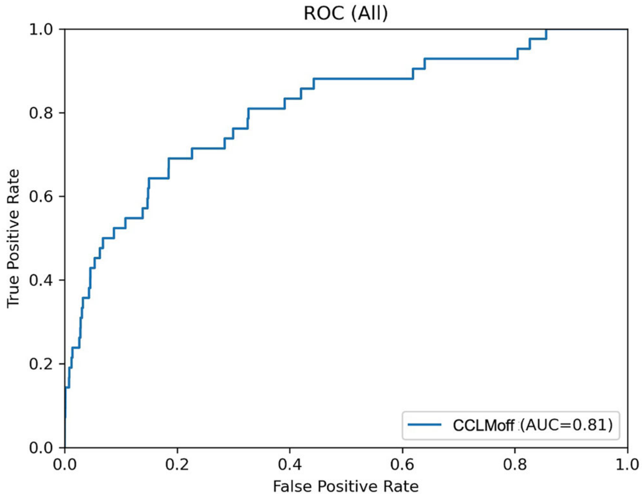

Fig. 4 | The CCLMoff performance on uncanonical length sgRNA (len = 19, 21). Despite being trained solely on canonical 20- nt sgRNAs, CCLMoff achieved an AUROC of 0.81 on this unseen dataset, highlighting its strong generalization ability. This result underscores the advantage of the underlying language model in handling variable-length inputs, which is crucial for real-world sgRNA design where non-canonical lengths are frequently employed to optimize CRISPR targeting efficiency.

CCLMoff achieved considerable performance on uncanonical length sgRNA

One of the notable advantages of the language model is its ability to handle variable-length inputs. This feature is particularly valuable in the context of sgRNA design, where several sgRNAs with non-canonical lengths are engineered to optimize cutting sites and on-target efficiency. Despite the need for such sgRNAs with uncanonical length, no existing in silico model can predict off-target effects for these uncanonical length sgRNAs. To address this gap, we trained CCLMoff on a dataset with sgRNAs of length 20 and evaluated its performance on datasets with sgRNAs of lengths 19 and 21. CCLMoff achieved an AUROC of 0.8123, demonstrating a considerable capability for off-target prediction (Fig. 4). This result reveals that CCLMoff, when trained on a 20nt sgRNA dataset, can also effectively depict the off-target landscape for sgRNAs of different lengths. Furthermore, the full version of CCLMoff, which will incorporate a comprehensive dataset encompassing various sgRNA lengths, is expected to offer even more robust generalization capabilities. By training on such diverse data, CCLMoff aims to develop a more universal model for off-target predictions, capable of accurately predicting off-target effects across a wider range of sgRNA lengths. This advancement will significantly enhance the flexibility and applicability of the model in genomic research, ensuring that it can be effectively used for designing sgRNAs with non-canonical lengths without sacrificing prediction accuracy. Thus, CCLMoff not only addresses current limitations but also sets the stage for more versatile and reliable off-target prediction tools in the future.

CCLMoff reveals the PAM-near region motif for off-target prediction

The attention map (Fig. 5) derived from CCLMoff reveals that the model places greater emphasis on the PAM-proximal region (positions 16–20), which aligns with the seed region identified in previous studies[26]. This consistency with established research highlights the accuracy and reliability of CCLMoff in capturing critical elements of the off-target occurrence mechanism. Interestingly, a similar pattern is observed on the target site, with a slight positional shift to positions 14–18. These findings suggest that CCLMoff effectively captures the off-target occurrence mechanism, demonstrating its capability to identify and emphasize key genomic regions involved in off-target effects. The model's alignment with known biological mechanisms further validates its effectiveness and potential for accurate off-target prediction in CRISPR applications.

Discussion

The development of genome editing technologies has opened new avenues for disease treatment, but the potential off-target effects, particularly in the widely used CRISPR/Cas9 system, remain a significant concern. Current off-target detection methods rely on high-throughput sequencing, which is both expensive and time-consuming, and are often limited in providing prior knowledge for effective sgRNA design. To overcome these limitations, several in silico models have been developed for off-target detection; however, many models still struggle with generalization ability. In this study, we present a deep learning framework for CRISPR/Cas9 off-target prediction, namely CCLMoff. Built on the most comprehensive dataset and pretrained language models to date, CCLMoff employs a two-step cascade strategy: off-target searching and off-target scoring. The first step identifies as many mismatch-based off-target candidates as possible, while the second step scores these candidates to determine which mismatches are tolerable for the CRISPR/Cas9 system. CCLMoff outperforms state-of-the-art models across various scenarios, including both cross-validation and external validation, demonstrating its ability to accurately identify off-target sites and capture

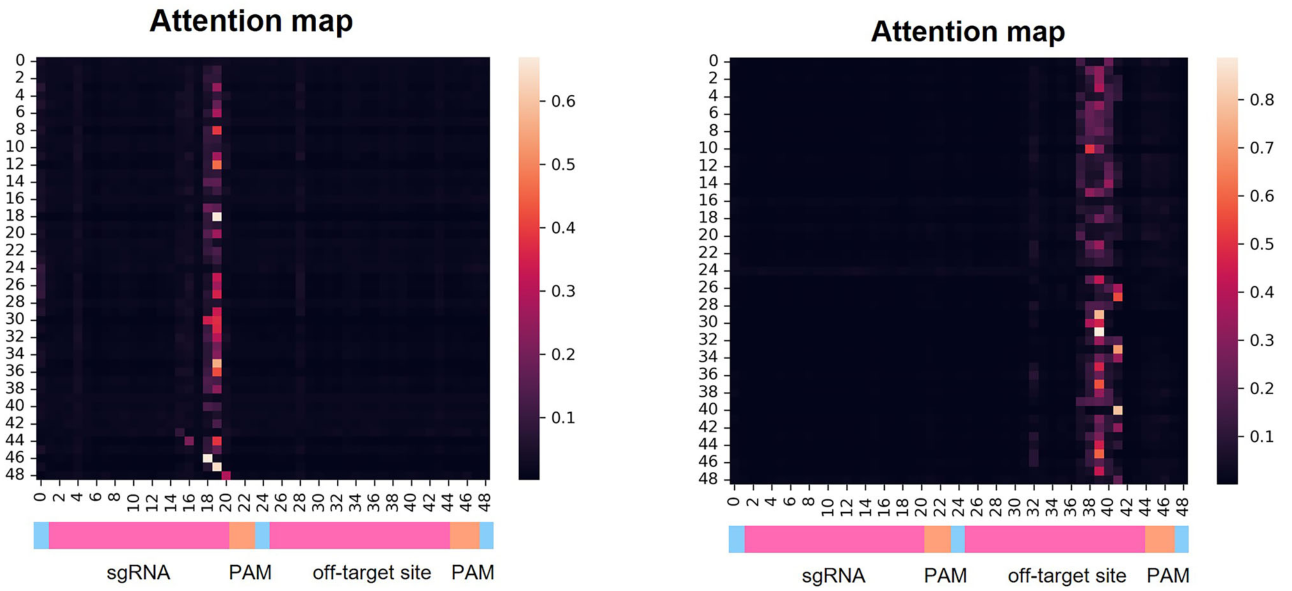

Fig. 5 | Model interpretation analysis for CCLMoff. The attention map of multi-head attention (layer=0, head=9 and layer=0, head=12) extracted from CCLMoff reveal that CCLMoff lay higher weight in the position 16–20 on sgRNA and position 14–18 on off-target site.

general off-target occurrence patterns. The current version of CCLMoff is a sequence-only model, making it user-friendly. Besides, CCLMoff achieved a comparable performance with the CCLMoff-Epi with the extra channel of four epigenetic information. Thus, CCLMoff lays the groundwork for optimizing sgRNA design and has the potential to accelerate advancements in genome editing technologies for therapeutic applications.

In this study, we demonstrate that increasing the amount of data enhances the model's ability to extract genomic patterns captured by the Cas9 protein. To this end, we compiled a comprehensive dataset utilizing various deep sequencing methods from multiple sources. Although we assembled the most extensive dataset available for constructing the CCLMoff model, it still falls short of fully leveraging the potential of a language model. To further improve the model's capabilities, efforts such as incorporating off-target datasets from other Cas proteins, like Cas12 and Cas13, could reveal similar off-target patterns that would enhance model development. In addition, integrating on-target efficiency datasets could provide valuable sgRNA-related information, enabling more effective feature extraction. Overall, the incorporation of diverse datasets related to the CRISPR/Cas system into the language model could lay the foundation for a versatile and robust model for the CRISPR/Cas system, capable of supporting a broad range of genome editing applications.

Existing in silico models primarily focus on identifying the contribution of mismatches to off-target effects, relying solely on sequence information for model construction. However, the CRISPR/Cas9 system involves a protein-nucleic acid interaction, with two key components: the sgRNA and the target sites. Recent research[49] has shown that the secondary structure of sgRNA plays a crucial role in enhancing CRISPR/Cas9 cleavage efficiency by introducing a structural lock into the hairpin structure. This finding suggests that incorporating structural information into model construction could further capture the true interaction dynamics of the sgRNA-target site duplex for off-target prediction, potentially improving model performance and providing deeper insights into the mechanisms of off-target occurrences. Moreover, several studies[26,40] have highlighted the importance of epigenetic modifications and chromatin status at the target site in predicting off-target effects. DeepCRISPR, for instance, employs additional channels to encode epigenetic information, including DNase, CTCF, H3K4me3, and RRBS. Similarly, CrisprSQL integrates epigenetic data such as CTCF, DNase, H3K4me3, RRBS, and DRIP, providing cell line-specific annotations that enhance the precision of off-target effect prediction. In addition, energy features generated by computational models can be incorporated as extra inputs to represent the mutual information between the sgRNA and the target site[50]. By incorporating structural information, epigenetic annotations, and energy features, future models can provide a more comprehensive and explicit framework for off-target prediction, improving both performance and understanding of the underlying mechanisms.

The current off-target prediction methods have been primarily utilized for two key purposes. The first is to evaluate the activity levels of a specific sgRNA on off-target regions, where CCLMoff has demonstrated strong performance. The second purpose is to assess the off-target effects of designed sgRNAs in advance. Previous models, such as Elevation-Aggregate[18] and CRISPR-Net-Aggregate[27], follow an aggregate strategy, where the potential sgRNA-target site pairs are summed into an overall off-target score for a designed sgRNA. In the future, CCLMoff could introduce a CCLMoff-Aggregate version that adopts a similar approach, enabling more effective evaluation of proposed sgRNAs. Predicting on-target efficiency is an important aspect of the CRISPR/Cas9 system, and recent studies have shown that deep learning frameworks such as DeepHF[51] and CRISPRon[52] have great potential for on-target efficiency prediction. However, these models have not yet utilized language models for this purpose. With the emergence of high-throughput sequencing-based on-target efficiency determination approaches, comprehensive on-target efficiency datasets can be built to enhance in silico sgRNA efficiency prediction. The current version of CCLMoff is equipped with a pretrained language model on a comprehensive sgRNA-target site pair dataset, which suggests that the model can also be applied to on-target efficiency prediction through transfer learning. Fine-tuning a CRISPR-based language model in this way may significantly improve sgRNA on-target efficiency prediction. Moreover, the language model has demonstrated its capability in multi-modality applications, which means that it can directly incorporate structural information about sgRNA, such as secondary structure and chromatin status at the cleavage site.

Recent studies have demonstrated that the editing outcome of a specific sgRNA is reproducible, suggesting that the editing outcome of a specific sgRNA can be predicted. Several models have been proposed for cleavage outcome prediction, such as CROTON[53], Lindel[54], and Forecast[55], all of which have shown promising performance. In future versions of CCLMoff, we plan to include a function for cleavage outcome prediction to provide a more comprehensive evaluation of sgRNA performance. To achieve this goal, we plan to integrate CCLMoff into the existing sgRNA design platform to provide a comprehensive evaluation of sgRNA on on-target efficiency, off-target effect, and cleavage outcome. By incorporating such a prediction tool, the researchers can better evaluate the potential outcomes of a specific sgRNA and design the most appropriate sgRNA for their research purposes. Moreover, we aim to develop a user-friendly interface for the platform, which enables users to input the target gene information and obtain the optimal sgRNA sequences based on the comprehensive evaluation. With such a platform, researchers can save a significant amount of time and resources and accelerate their research in the field of genome editing.

Data availability

The comprehensive curated dataset is available on Figshare[56] with the https://doi.org/10.6084/m9.figshare.27080566.v2. All other data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The source code for CCLMoff is available at https://github.com/duwa2/CCLMoffand Zenodo[57]. The model of AttenToCrispr is available at https://github.com/qiaoliuhub/AttnToCrispr. The CNN_std is available at https://github.com/MichaelLinn/off_target_prediction. CRISPR-Net is available at https://github.com/JasonLinjc/CRISPR-Net. CRISPRoff is available at https://github.com/RTH-tools/crisproff. CCTop is available at https://cctop.cos.uni-heidelberg.de/.

Received: 21 September 2024; Accepted: 22 May 2025;
Published online: 06 June 2025

References

1. Oakes, B. L. et al. CRISPR-Cas9 circular permutants as programmable scaffolds for genome modification. *Cell* **176**, 254–267 (2019).
2. Abudayyeh, O. O. et al. RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017).
3. Yuan, J. et al. Genetic modulation of RNA splicing with a CRISPR-guided cytidine deaminase. *Mol. Cell* **72**, 380–394 (2018).
4. Papasavva, P., Kleanthous, M. & Lederer, C. W. Rare opportunities: CRISPR/Cas-based therapy development for rare genetic diseases. *Mol. Diagn. Ther.* **23**, 201–222 (2019).
5. Ousterout, D. G. et al. Reading frame correction by targeted genome editing restores dystrophin expression in cells from Duchenne muscular dystrophy patients. *Mol. Ther.* **21**, 1718–1726 (2013).
6. Kennedy, E. M. & Cullen, B. R. Gene editing: a new tool for viral disease. *Annu. Rev. Med.* **68**, 401–411 (2017).
7. Xiong, X., Chen, M., Lim, W. A., Zhao, D. & Qi, L. S. CRISPR/Cas9 for human genome engineering and disease research. *Annu. Rev. Genomics Hum. Genet.* **17**, 131–154 (2016).
8. Huang, C.-H., Lee, K.-C. & Doudna, J. A. Applications of CRISPR-Cas enzymes in cancer therapeutics and detection. *Trends Cancer* **4**, 499–512 (2018).
9. Chavez, M., Chen, X., Finn, P. B. & Qi, L. S. Advances in CRISPR therapeutics. *Nat. Rev. Nephrol.* **19**, 9–22 (2023).
10. Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target effects in CRISPR/Cas9-mediated genome engineering. *Mol. Ther. Nucleic Acids* **4**, e264 (2015).
11. Kwon, J. et al. Extru-seq: a method for predicting genome-wide Cas9 off-target sites with advantages of both cell-based and in vitro approaches. *Genome Biol.* **24**, 1–20 (2023).
12. Zhang, L. et al. Systematic in vitro profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes. *Nucleic Acids Res.* **48**, 5037–5053 (2020).
13. Kim, D. et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243 (2015).
14. Tsai, S. Q. et al. Circle-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).
15. Wienert, B. et al. Unbiased detection of CRISPR off-targets in vivo using discover-seq. *Science* **364**, 286–289 (2019).
16. Ortinski, P. I., O'Donovan, B., Dong, X. & Kantor, B. Integrase-deficient lentiviral vector as an all-in-one platform for highly efficient CRISPR/Cas9-mediated gene editing. *Mol. Ther. Methods Clin. Dev.* **5**, 153–164 (2017).
17. Tsai, S. Q. et al. Guide-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
18. Listgarten, J. et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. Biomed. Eng.* **2**, 38–47 (2018).
19. Bao, X. R., Pan, Y., Lee, C. M., Davis, T. H. & Bao, G. Tools for experimental and computational analyses of off-target editing by programmable nucleases. *Nat. Protoc.* **16**, 10–26 (2021).
20. Bae, S., Park, J. & Kim, J.-S. Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
21. Labun, K. et al. Chopchop v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).
22. Bradford, J. & Perrin, D. A benchmark of computational CRISPR-Cas9 guide design methods. *PLoS Comput. Biol.* **15**, e1007274 (2019).
23. Dobson, L., Reményi, I. & Tusnády, G. E. Cctop: a consensus constrained topology prediction web server. *Nucleic Acids Res.* **43**, W408–W412 (2015).
24. Cao, Q. et al. Crispr-focus: a web server for designing focused crispr screening experiments. *PLoS ONE* **12**, e0184281 (2017).
25. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biol.* **19**, 1–13 (2018).
26. Chuai, G. et al. Deepcrispr: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 1–18 (2018).
27. Lin, J., Zhang, Z., Zhang, S., Chen, J. & Wong, K.-C. Crispr-net: a recurrent convolutional network quantifies CRISPR off-target activities with mismatches and indels. *Adv. Sci.* **7**, 1903562 (2020).
28. Yan, J. et al. Benchmarking and integrating genome-wide CRISPR off-target detection and prediction. *Nucleic Acids Res.* **48**, 11370–11379 (2020).
29. Cameron, P. et al. Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nat. Methods* **14**, 600–606 (2017).
30. Zou, R. S. et al. Improving the sensitivity of in vivo CRISPR off-target detection with discover-seq+. *Nat. Methods* **20**, 706–713 (2023).
31. Lazzarotto, C. R. et al. Change-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity. *Nat. Biotechnol.* **38**, 1317–1327 (2020).
32. Tsai, S. Q. & Joung, J. K. Defining and improving the genome-wide specificities of CRISPR–Cas9 nucleases. *Nat. Rev. Genet.* **17**, 300–312 (2016).
33. Kim, D. & Kim, J.-S. Dig-seq: a genome-wide CRISPR off-target profiling method using chromatin DNA. *Genome Res.* **28**, 1894–1900 (2018).
34. Wang, X. et al. Unbiased detection of off-target cleavage by CRISPR-Cas9 and talens using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **33**, 175–178 (2015).
35. Frock, R. L. et al. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–186 (2015).
36. Pan, X. et al. Massively targeted evaluation of therapeutic CRISPR off-targets in cells. *Nat. Commun.* **13**, 4049 (2022).
37. Shen, T. et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*. pp. 1–12 (Nature Publishing Group US New York, 2024).

38. Consortium, T. R. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **49**, D212–D220 (2021).
39. Guan, Z. & Jiang, Z. Transformer-based anti-noise models for CRISPR-Cas9 off-target activities prediction. *Brief. Bioinforma.* **24**, bbad127 (2023).
40. Störtz, F. & Minary, P. crisprsql: a novel database platform for CRISPR/Cas off-target cleavage assays. *Nucleic Acids Res.* **49**, D855–D861 (2021).
41. Sherkatghanad, Z., Abdar, M., Charlier, J. & Makarenkov, V. Using traditional machine learning and deep learning methods for on-and off-target prediction in CRISPR/Cas9: a review. *Brief. Bioinforma.* **24**, bbad131 (2023).
42. Liu, Q., He, D. & Xie, L. Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas system using attention boosted deep learning and network-based gene feature. *PLoS Comput. Biol.* **15**, e1007480 (2019).
43. Zhang, Z.-R. & Jiang, Z.-R. Effective use of sequence information to predict CRISPR-Cas9 off-target. *Comput. Struct. Biotechnol. J.* **20**, 650–661 (2022).
44. Fu, R. et al. Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity. *Nat. Commun.* **13**, 474 (2022).
45. Nguyen, E. et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Adv. Neural Inf. Process. Syst.* **36**, 43177–43201 (2023).
46. Luo, Y., Chen, Y., Xie, H., Zhu, W. & Zhang, G. Interpretable CRISPR/Cas9 off-target activities with mismatches and indels prediction using bert. *Comput. Biol. Med.* **169**, 107932 (2024).
47. Lin, J. & Wong, K.-C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics* **34**, i656–i663 (2018).
48. Carlson-Stevermer, J. et al. Crispoff enables spatio-temporal control of CRISPR editing. *Nat. Commun.* **11**, 5041 (2020).
49. Riesenberg, S., Helmbrecht, N., Kanis, P., Maricic, T. & Pääbo, S. Improved grna secondary structures allow editing of target sites resistant to CRISPR-Cas9 cleavage. *Nat. Commun.* **13**, 489 (2022).
50. Mathis, N. et al. Predicting prime editing efficiency and product purity by deep learning. *Nat. Biotechnol.* **41**, 1151–1159 (2023).
51. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
52. Xiang, X. et al. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat. Commun.* **12**, 3238 (2021).
53. Li, V. R., Zhang, Z. & Troyanskaya, O. G. Croton: an automated and variant-aware deep learning framework for predicting CRISPR/Cas9 editing outcomes. *Bioinformatics* **37**, i342–i348 (2021).
54. Chen, W. et al. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Res.* **47**, 7989–8003 (2019).
55. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* **37**, 64–72 (2019).
56. Du, W. et al. CCLMoff: A CRISPR/Cas9 System Off-target Prediction Tool Using Language Model. https://doi.org/10.6084/m9.figshare.27080566.v2 (2024).
57. Du, W. et al. CCLMoff: a CRISPR/Cas9 system off-target prediction tool using language model. *Zenodo* https://doi.org/10.5281/zenodo.15385508 (2025).
58. Anderson, K. R. et al. CRISPR off-target analysis in genetically engineered rats and mice. *Nat. Methods* **15**, 512–514 (2018).
59. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
60. Chen, J. S. et al. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
61. Ran, F. A. et al. In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
62. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).

Acknowledgements

This work was jointly supported by the National Key R&D Plan of China (2021YFA0805901), the National Natural Science Foundation of China (82070199), the Guangdong Basic and Applied Basic Research Fund (2021A1515220078), the Guangdong Provincial R&D Project in Key Areas (2022B0202110002), and the Special Project in Key Fields of Guangdong Provincial Universities (2021ZDZX2050).

Author contributions

L.Z., X.Z. and D.T. conceived the project. W.D. and L.Z. designed the methodology and conducted the experiments. L.Z., X.Z. and D.T. supervised the study. W.D., Y.Z., K.D., Z.Z. and Q.Y. drafted the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08275-6.

Correspondence and requests for materials should be addressed to Liang Zhao, Xiangxing Zhu or Dongsheng Tang.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Aylin Bircan.