Sequence analysis

AttCRISPR : a spacetime interpretable model for prediction of sgRNA on-target activity

Liming Xiao [1], Yunqi Wan [1] and Zhenran Jiang [1,*]

[1] School of Computer Science & Technology, East China Normal University, Shanghai, 200062, China.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Abstract

Motivation: More and more Cas9 variants with higher specificities are developed to avoid the off-target effect, which bring a significant volume of experimental data. Conventional machine learning performs poorly on these datasets, while the methods based on deep learning are often lack of interpretability, which makes researchers have to trade-off accuracy and interpretability. It is necessary to develop a method which can not only match deep learning based methods in performance, but also with strong interpretability that can be comparable to conventional machine learning methods.

Results: To overcome these problems, we design and implement AttCRISPR, an intrinsic interpretable method based on deep learning to predict the on-target activity. The advantage of AttCRISPR lies in using the ensemble learning strategy to stack available encoded-based methods and embedded-based methods with strong interpretability. Comparison with the state-of-the-art methods (WT-SpCas9, eSpCas9(1.1), SpCas9-HF1), AttCRISPR can achieve an average spearman value of 0.872, 0.867, 0.867, respectively on several public datasets, which is superior to these methods. Furthermore, benefits from two attention modules, spatial and temporal attention modules, AttCRISPR has strong interpretability. Through them, we can understand the decisions made by AttCRISPR at both global and local levels without other post hoc explanations techniques. With the trained model, we reveal the preference for each position-dependent nucleotide on the sgRNA (short guide RNA) sequence in each dataset at global level. And at local level, we prove that interpretability of AttCRISPR can be used to guide researcher to design sgRNA with higher activity.

Availability: The example code are available at https://github.com/South-Walker/AttCRISPR

Contact: zrjiang@cs.ecnu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) / CRIPSR associated protein 9 (Cas9) systems are preferred over other biological research and human medicine technologies now, beacuse of it's efficiency, robustness and programmability. Cas9 nucleases can be directed by sgRNAs to introduce site-specific DNA double-stranded breaks in target, so to enable editing site-specific regions within genome (Jinek *et al.*, 2012; Cong *et al.*, 2013; Mali *et al.*, 2013). CRISPR/Cas9, to a large extent, has developed genetic therapies at the cellular level, while there are still severe medical disadvantages even now which has greatly hindered the further clinical application of the CRISPR/Cas9 systems. One of these disadvantages is due to unexpected insertion and deletion caused by off-target effect (Rubeis and Steger, 2018; Kang *et al.*, 2016; Ishii, 2017; Liang *et al.*, 2015). To overcome this disadvantage, a solution is to engineer CRISPR / Cas9 with higher specificities. That's why more and more higher specificities Cas9 variants, such as enhanced SpCas9 (eSpCas9(1.1)), Cas9-High Fidelity (SpCas9-HF1) (Slaymaker *et al.*, 2016; Ishii, 2017), hyper-accurate Cas9 (HypaCas9) (Kleinstiver *et al.*, 2016), been developed and bring a significant volume of experimental data, which means that researchers have to face the difficulty of analyzing such huge and heterogeneous data. The activity of chosen sgRNA sequence determines the efficiency of genome editing, This fact indicates that it's meaningful to explore an efficient approach to predict sgRNA activity and even guide sgRNA design.

1
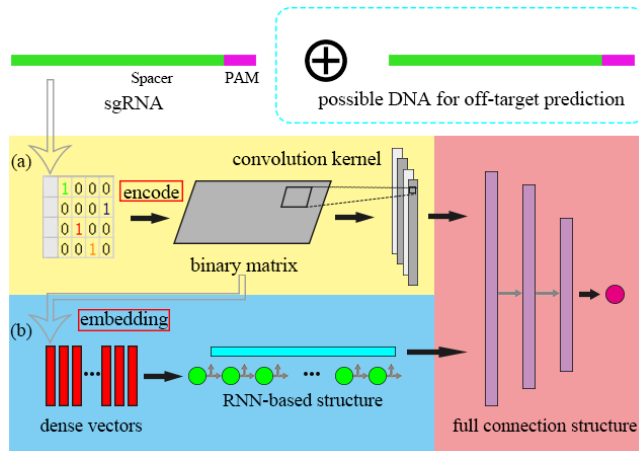
Fig. 1. Two categories of the deep learning models used in sgRNA related task. (a) Model work in spatial domain. In spatial domain, the base sequence is encoded into a binary matrix (or a binary image). Since convolution has great advantages in extracting spatial features, the CNN is an excellent tool in spatial domain. (b) Model work in temporal domain. In temporal domain, the base sequence (represented by the binary matrix) is embedded into a sequence of high-dimensional vector, in which the RNN perform better. In addition, we note that the last layers of neural network are usually full connection structure (not necessarily), which greatly increases the difficulty of understanding the decisions of model.

In practice, there have been a number of applications and toolkits applied in this task. In the earlier studies, the methods *in silico* are categorized into three types: (1) alignment-based, (2) hypothesis-driven and (3) learning-based (Chuai *et al.*, 2018). Recently, we noticed that the last type of method seems to be getting more attention because of huger and huger datasets (Liu *et al.*, 2019a). Learning-based method is essentially a computational model built by machine learning algorithms, not only conventional machine learning but also deep learning. Some studies on HT_ABE and HT_CBE (two gene editing tools grew out of CRISPR) have shown that deep learning based models often outperformed conventional machine learning, when the number of sgRNAs in the dataset reached a certain level (Kim *et al.*, 2018, 2019; Song *et al.*, 2020). Per contra, conventional machine learning algorithms, such as linear regression, logistic regression and the decision tree, are often more interpretable due to the fewer parameters and clearer mathematical assumptions. In short, what was needed for developer is to trade-off accuracy and interpretability. Muhammad Rafid *et al.* consider deep-learning models as black boxes and believe they lack interpretability, motivated by the empirical assertion, they turn to build a model based on conventional machine learning to compete with state-of-the-art deep learning models (Muhammad Rafid *et al.*, 2020). On the other hand, input perturbation based feature importance analysis become a preferred components to reveal the importance of features in deep learning models. Liu *et al.* use a sliding window of length 2 to extract dimeric as input and rank the position of dimeric by contribution to final output. One regret is that subject to the processing of the input sgRNA sequence, their analysis can not exactly on the nucleotide class. Further, SHAP, one of the most prominent of model explain techniques, has been widely used to understand the decision made by the model. Wang *et al.* develope DeepHF, a deep learning based model, and use Deep SHAP to revealed nucleotide contributions. In our understanding, the method based on input perturbation often requires better generalization ability of the model (even for artificial ridiculous noise data).

In addition, the interpretability of existing models are all at global level, and result is a general pattern in the datasets. For example, which position has a great impact on the final decision of the model and which position-dependent nucleotide has a positive impact on the activity. Lack of analysis at local level. For instance, which structure causes the low activity of a certain nucleotide sequence and how to improve its activity with a few modification. In light of the above, we believe it is critical to develop an effective model which can not only have good performance, but also with strong interpretability.

Deep neural network has shown its power in the study of CRISPR/Cas9 and its improved systems (Liu *et al.*, 2019a). Most of the deep neural networks existing are the combination of recurrent neural network (RNN), convolutional neural network (CNN), fully connected neural network (FNN), and their variants. As Figure 1 show, we found that the deep learning models used in sgRNA on-target activity (even for off-target effect) prediction tasks in recent years can be divided into the following two categories according to the encoding approach of the sgRNA sequence (sgRNA-DNA sequence pair, for off-target effect prediction):

1. Methods in spatial domain. Some previous studies have used the methods based CNN to predict sgRNA on-target activity or off-target effect (Chuai *et al.*, 2018; Kim *et al.*, 2018; Lin and Wong, 2018). They process sgRNA base sequence inputs with the help of one-hot encoding idea. In other words, they regard it as two-dimensional image data, and use convolution layer to extract potential features in spatial domain, It is worth noting that Zhang *et al.* adds bidirectional gated recurrent unit (BGRU, in short), a RNN variant, after pooling layer of classic CNN network (Zhang *et al.*, 2020). Our explanation is that BGRU assists CNN to extract spatial features in one dimension, under this belief it belong to this category.

2. Methods in temporal domain. These methods are not used for on-target activity or off-target effect prediction, until recently (Liu *et al.*, 2019b; Wang *et al.*, 2019; Liu *et al.*, 2020). They consider the nucleotides (can also dimer or polymer) in the sgRNA sequence as a word, then a trainable matrix (could be either supervised or unsupervised) is used to project the word to the dense real-valued space. This technology is called embedding, which generates the base embedding. However, base embedding is not spatially interpretable (different from one-hot encode). Almost all of the methods in temporal domain used in sgRNA on-target activity or off-target effect flatten the hidden state vector into a one-dimensional vector as the input of the fully connected layer. It is a pity that the temporal sequential dependency of hidden state vector are rarely noticed.

Attention mechanism has demonstrated its power in NLP, Statistical Learning, Speech and Computer Vision. It makes model tends to focus selectively on parts of the input, which is help in performing the task effectively. Previous observation have shown that Cas9 preferentially binds sgRNAs containing purines but not pyrimidines (Wang *et al.*, 2014) and multiple thymine in the spacer impairing sgRNA activity (Wu *et al.*, 2014), that is to say some specific nucleotides and base position need more attention compared to others. The above is the premise of introducing attention mechanism. Strictly speaking, we are not the first to bring attention mechanisms into this field. The most similar approach to ours is the work based on transformer by Liu *et al.*. They use transformer, a components based on attention mechanism, instead of RNN to improve the ability of temporal feature extraction, hence, enhance the performance of their model (Vaswani *et al.*, 2017; Liu *et al.*, 2019b). In our work, the interpretability benefit from attention mechanism is more focused. Our main contributions are as follows:

- Present a novel deep-learning model, which can extract potential feature representation of sgRNA sequence in both spatial and temporal domain parallelly. Finally, the ensemble learning method is used to combine the two to achieve better performance than current state-of-the-art models.

- Introduce attention mechanism into our model. As a result, it does not need post hoc explanations techniques based on input perturbation to explain itself. It is intrinsic interpretable in both temporal and spatial domains. In the spatial domain it's at global level, while at local level in the temporal domain.
- Throught ablation analysis and testing a series of possible network structures, we find there are multiple components and strategies can improve the performance of AttCRISPR, which could outperform current state-of-the-art tools on DeepHF dataset.

2 Materials and methods

2.1 Datasets

The dataset we used for training, validation and testing is built by Wang *et al.*. We extracted 55604, 58617, 56888 sgRNAs with activity (represented by insertion/deletion (indel)) for WT-SpCas9, eSpCas9(1.1) and SpCas9-HF1, respectively, from its source data.

2.2 Sequence encoding and embedding

For encoding process, we use the complementary base to represent the original base in sgRNA. Further, we use one-hot encode strategy, that is to say, we encoding each base in sgRNA into a four-dimensional vector (encode A,T,G,C into [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1], respectively), called one-hot vector. Then a sgRNA can be considered as a matrix $X_{oh} \in \mathbb{R}^{l \times 4}$, named one-hot matrix (a little sparse, since an one-hot vector is zero in all but one dimension). We believe it is meaningful to regard X_{oh} as a binary image, therefore, it is used as an input of CNN, which performs well in the image field. Meanwhile, as mentioned above, one-hot matrix is a little sparse.

To facilitate the training process, we can mapping each one-hot vector into a dense real-valued high-dimensional space, which is called embedding. In summarized, at the matrix level, the formula is as follows

$$X_e = X_{oh} E_m \tag{1}$$

where X_e named embedding matrix, $E_m \in \mathbb{R}^{4 \times m}$ is a trainable transformational matrix, m refers to the dimension of embedding space. We believe it is also meaningful to regard nucleotides in the sgRNA sequence as word, and the sgRNA sequence itself as a sentence, Guided by this belief E_m is the word embedding matrix and X_e is the sentence embedding in NLP. Therefore, X_e is used as an input of RNN (or its variant), which performs well in the NLP field.

Due to each element of X_{oh} is interpretable (representing whether there is a corresponding nucleotide type at the corresponding location), we call X_{oh} the spatial input, and the CNN work on X_{oh} is the method in spatial domain. In other hand, different from X_{oh}, X_e can only be explained in the first dimension (representing the embedding vector of corresponding nucleotide type), and embedding vector is difficult for humans to understand. That's why we call X_e the temporal input, and the RNN (or its variant) work on X_e belongs to the method in temporal domain.

2.3 Neural network architecture

Based on the categorization above, we assume that the method in spatial domain and in temporal domain are heterogeneous, which can satisfy the diversity premise of ensemble learning. Based on the assumption above and ensemble learning, we follow the stacking strategy to develop AttCRISPR which can extract potential feature representation of sgRNA sequence in both spatial and temporal domain parallelly. Further, we apply attention mechanisms in both spatial and temporal domain to enhance the interpretability of AttCRISPR.
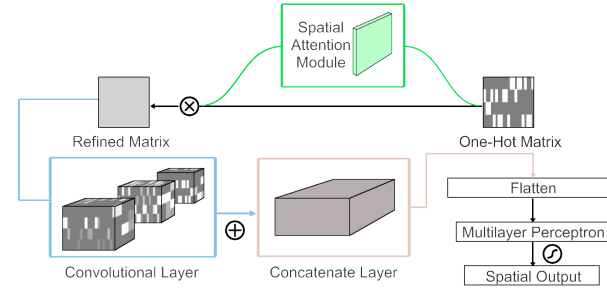
Fig. 2. The architecture of spatial domain method in AttCRISPR. The input of the method is encoded sgRNA sequence X_{oh}, a 21×4 one-hot matrix. Then refine it through a spatial attention module, which could tell us the importance of a specific matrix element (or just say, pixel). A simple CNN followed is applied to extract potential feature representation of sgRNA sequence. In the last step, we flatten the output of CNN network into a one dimensional vector and use a multilayer perceptron with sigmoid activation function to achieve the spatial output y_s.

2.3.1 First-order preference and second-order preference

In order to introduce the neural network architecture of AttCRISPR, Let's define first-order preference and second-order preference for convenience. Taking a simple linear regression model as an example, for input $X \in \mathbb{R}^l$, where l refers to the length of base sequences, predicted activity y is as following

$$y = AX \tag{2}$$

where $A \in \mathbb{R}^d$, the total differential of y in Equation (2) is as following

$$dy = \sum_i^d A_i dX_i \tag{3}$$

where A_i and X_i denotes the i-th dimension of the vector X and A, A_i indicates how dramatically the function changes as X_i changes in a neighborhood of X, in other words, the importance of X_i. That's why we'll call A first-order preference in our paper. Specifically, we use a vector A_i to build the first-order original preference at position i within sgRNA sequence, and X_i is an embeddedness of the i-th feature, then A and X are two matrix. Further, the final result can be weighted by a trainable non-negative weight vector $W \in \mathbb{R}^l$, as follow

$$y = W \cdot AX^T \tag{4}$$

then we define $\tilde{A}$ as the first-order combine preference matrix (or just first-order preference), which means $\tilde{A}$ can be expressed linearly by A as follow

$$\tilde{A} = BA \tag{5}$$

where the weight matrix $B \in \mathbb{R}^{l \times l}$ is learned through attention mechanism, which we'll call the second-order preference matrix in our paper due to its calculation is based on first-order preference, it can explain how a particular pattern containing two nucleotides affects the base sequence. Then the predicted value can be expressed as

$$y = W \cdot \tilde{A} X_e^T \tag{6}$$

2.3.2 Method in spatial domain

As Figure 2 demonstrated, the method in spatial domain relies on the CNN. As previously mentioned, sgRNA sequence has been encoded into a 21×4 one-hot matrix X_{oh}, and we regard X_{oh} as a binary image. Then, convolution kernels with different size are used to extract potential spatial features just like what other do in computer vision. According to
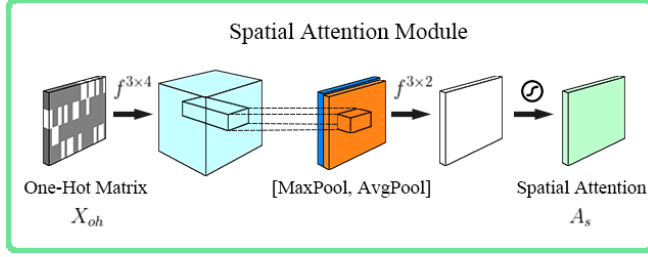
Fig. 3. Details of spatial attention module. A convolution layer is used to generate multi-channel map from X_{oh}. Then concatenated the output of both max-pooling and average-pooling method and forward it to the last convolution layer. A sigmoid function is used to map the final result to a range of zero to one at last, which generates the spatial first-order preference matrix A_s.

the foregoing, the spatial attention module proposed by Woo *et al.* can be applied in our method, which has been used to improve the performance of CNN in vision tasks.

As shown in Figure 3, given an one-hot map X_{oh}, spatial attention module generating spatial attention matrix $A_s \in \mathbb{R}^{l \times 4}$ with the same as X_{oh} in shape. Each element of A_s is constrained to a range of zero to one, implemented by a sigmoid function, which reflects the importance of the corresponding elements of X_{oh}. The overall spatial attention process can be summarized as:

$$\begin{cases} X_{mc} = f^{3 \times 4}(X_{oh}) \\ A_s = \sigma(f^{3 \times 2}([AvgPool(X_{mc}); MaxPool(X_{mc})])) \\ X_{rf} = A_s \otimes X_{oh} \end{cases} \quad (7)$$

where $f^{p \times q}$ represents a convolution operation with the filter size of $p \times q$, $p, q \in \mathbb{Z}^+$, X_{mc} is a multi-channel map generated by X_{oh}, σ denotes the sigmoid function, AvgPool denotes the average-pooling operation, MaxPool denotes the max-pooling operation, $\otimes$ denotes element-wise multiplication. Spatial attention matrix A_s formally conforms to our proposed definition of first-order preference (each element of X_{oh} is multiplied by the corresponding element of A_s), in other words, element of A_s reveal how important the corresponding elements in X_{oh} is. We think it can reveal the preference of the scoring function at each position. For instance, follow the encoding rules above, we train spatial domain part of AttCRISPR with the WT-SpCas9 dataset. Then take the average of all spatial attention matrix, and the element in the first row and third column are closer to 1, which means when calculate the final score, G typically may have a important contribution at first position within sgRNA sequence. In fact, this corresponds to some early studies concerning the Human (hU6) promoter, which is believed to require G as the first nucleotide of its transcript (Jinek *et al.*, 2012; Cong *et al.*, 2013; Mali *et al.*, 2013).

2.3.3 Method in temporal domain

As Figure 4 show, temporal domain part of AttCRISPR relies on the RNN (or its variant). As previously mentioned, we mapping each one-hot vector into a dense real-valued high-dimensional space follow the Equation (1), which generates the embedded matrix X_e. And we regard X_e as a sequential data, or temporal data. RNN (or its variant) has showed outstanding performance in the tasks with temporal data (for instance, NLP, sequential recommendation). That's why we prefer to use it to extract potential temporal features. To be precise, we prefer the architecture of encoder-decoder which has been proven to be effective in Seq2Seq task. Two main differences we have to face are that sgRNA is not a natural language in the traditional sense, and we don't have to translate it to other sequence. To accommodate them, the embedded matrix X_e is used as input of both the encoder and decoder, and the sequence of decoder is
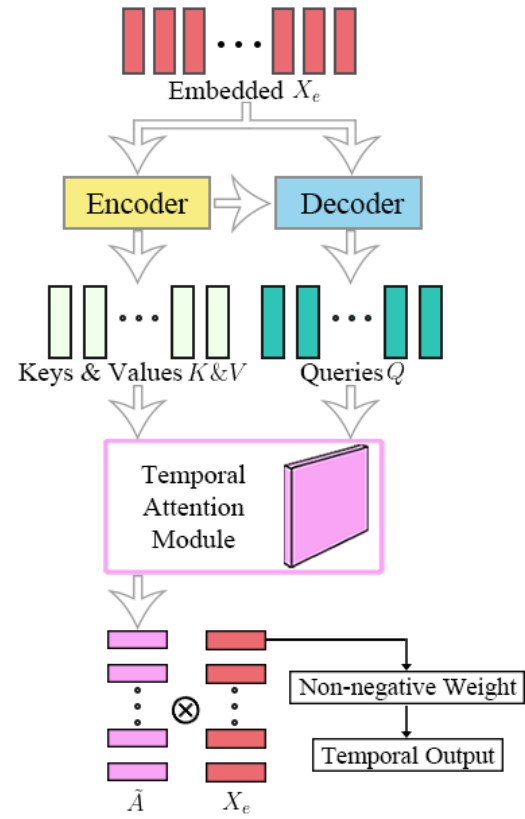
Fig. 4. The architecture of temporal domain method in AttCRISPR. The input of the method is embedding sgRNA sequence X_e. Then Keys K, Values V and Queries Q is generated through a classic encoder-decoder structure which is need for temporal attention module. Next, the temporal attention module generates the first-order preference $\tilde{A}$. Each of the row vectors in matrix $\tilde{A}$ represents the base preference of sgRNA at the corresponding position, we use their dot product with the corresponding row vector in embedded X_e to build the score of corresponding position. Hence, a full connection layer is used to weighted average them and achieve the temporal output y_t.

to build the first-order preference of sgRNA sequence $\tilde{A}$. As mentioned above, the predicted value y should satisfy Equation (6).

On this basis, we apply the idea of attention mechanism which has been widely used in NLP tasks to AttCRISPR in the method of temporal domain, and name it temporal attention module (Luong *et al.*, 2015; Vaswani *et al.*, 2017). According to the idea of Vaswani *et al.*, temporal attention module
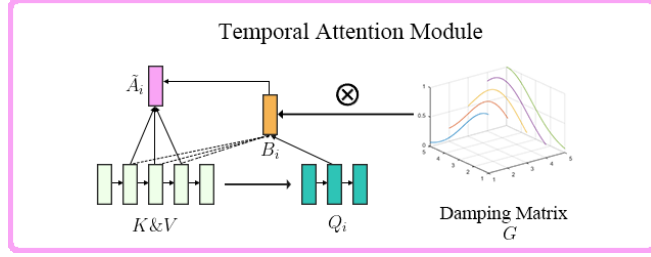
Fig. 5. Details of temporal attention module. To generate the first order preference vector in the i position of sgRNA $\tilde{A}_i$. First, the i-th row vector of the querys matrix Q_i is multiplied by the transpose of the keys mateix K^T, and apply a softmax function to obtain a preliminary weights vector on the values. Second, to favor the alignment points near i, the weights vector obtained is multiplied element-by-element by the i-th row vector of the damping matrix G, as Equation (11) has shown. G_{ij} can be regarded as the result of place a Gaussian distribution centered around i, then sampling the position j (a scaling factor is used to ensure the sum of G_I is l). Then we achieve the second-order preference matrix B, it is also a weights matrix on the values. So the first-order preference matrix $\tilde{A}$ come from the product of B and values matrix V. Temporal attention module generates the temporal first-order preference matrix $\tilde{A}$ and the temporal second-order preference matrix B.

satisfies the following equation

$$Attention(Q, K, V) = align(Q, K)V \quad (8)$$

where the $align(\cdot)$ is put forward by Luong *et al.*. Q, K, V are queries, keys and values matrix accordingly in the paper of Vaswani *et al.*.

As Figure 5 show, in our attention module they are calculated by the following equation

$$\begin{cases} K_i = Encoder(X_{ei}, \theta_E, K_{i-1}) \\ Q_i = Decoder(X_{ei}, \theta_D, Q_{i-1}) \\ V = K \end{cases} \quad (9)$$

where vector K_i, Q_i denotes the i-th row of the matrix K and Q accordingly, $Encoder(\cdot)$ and $Decoder(\cdot)$ are independent GRU units, θ_E and θ_D denote all the related parameters of GRU networks accordingly. In the actual implementation, we apply the bidirectional GRU networks for better performance, and for the sake of conciseness, we show a conventional GRU network here. The function $align(\cdot)$ is as follows

$$B = align(Q, K) \quad (10)$$

$$B_i = softmax(Q_i K^T) \otimes G_i \quad (11)$$

$$G_{ij} = \begin{cases} exp(\frac{(i-j)^2}{-2\sigma}) & , |i-j| \leqslant \sigma \\ 0 & , |i-j| > \sigma \end{cases} \quad (12)$$

where the matrix $B \in \mathbb{R}^{l \times l}$ is the second-order preference we need, and vector B_i denotes the i-th row of the matrix B. $G \in \mathbb{R}^{l \times l}$ is the damping matrix base on Gaussian function. Since a simple belief that the closer the base is to the i-th position, the more it affects the i-th position normally, we use the damping matrix G to constrain the network learning. σ represents a threshold of length, any base over this length from the position i is not considered to be affected. Further, if we think of the values matrix as a vector form of the first-order preference A in Equation (5), we can reach the following equation

$$\tilde{A} = BV \quad (13)$$

according to above, values matrix V comes from the hidden states of a bidirectional GRU networks, which is usually hard to understand. While B is the second-order preference matrix obtained by the attention mechanism. We believe that the j-th dimension of B_i, denoted as B_{ij}, can reveal the effect of the base at position j on position i in the biological sense.

2.3.4 Ensemble model following stacking strategy

Some indirect sgRNA features, which can't be obtained directly by deep learning, including position accessibilities of secondary structure, stem–loop of secondary structure, melting temperature, and GC content are strongly associated with sgRNA activity (Wang *et al.*, 2016, 2019). It's worth noting that in the work of Wang *et al.*, hand-crafted biological features didn't be normalized. Since the wide range of data distribution, we normalize it based on the standard score (Z-Score).

Then we use a simple full connection network to extract the indirect features, and call the output of fully connection network y_{bio}. As mentioned above, we assume that the method in spatial domain and in temporal domain can satisfy the diversity premise of ensemble learning. That's why we follow the stacking strategy, to integrate the methods in time domain and space domain. Specifically, the y_{bio}, the spatial output y_s and the temporal output y_t we got earlier are concatenated and then weighted averaging is performed through a full connection layer as follow

$$y = W[y_{bio}; y_s; y_t] \quad (14)$$

where, y is the final prediction value of AttCRISPR, W is the weight learned by the full connection network. In the actual implementation, we freeze the network in the spatial domain and temporal domain firstly, in order to make our network focused on learning the weight W. Then the parameters of the entire network are adjusted in the fine tuning of AttCRISPR.

2.4 Current prediction methods

As far as we know, DeepHF, the state-of-the-art method, achieved Spearman correlation coefficients of 0.867, 0.862 and 0.860 for WT-SpCas9, eSpCas9(1.1) and SpCas9-HF1, respectively (Wang *et al.*, 2019). Another tool worth noting is CRISPRpred(SEQ), which is based on conventional machine learning and need complicated human feature engineering exercise (Muhammad Rafid *et al.*, 2020). They also test their tool with this dataset and achieved Spearman correlation coefficients of 0.838, 0.830 and 0.821 for WT-SpCas9, eSpCas9(1.1) and SpCas9-HF1, respectively (without any hyperparameter tuning).

In order to make the comparison apples to apples, we follow the same strategy as Wang *et al.* to design our experiments. To be more specific, each set is shuffled and divided into three parts, 76.5%, 15% and 8.5% of the relevant data was used as the training, test and validation set respectively in

Table 1. The methods compared in our work and brief description of these methods

Method	Neural	End2end	Description
CNN*	Yes	Yes	Naive CNN
RNN*	Yes	Yes	Bidirectional long short-term memory neural network
XGBoost*	No	Yes	Extreme Gradient Boosting regression tree
MLP*	Yes	Yes	Multilayer perceptron
DeepHF*	Yes	No	Bidirectional long short-term memory neural network (with hand-crafted biological features)
CRISPRpred(SEQ)#	No	No	A conventional machine learning pipeline
SpAC	Yes	Yes	Spatial AttCRISPR
TAC	Yes	Yes	Temporal AttCRISPR
EnAC	Yes	Yes	Ensemble AttCRISPR (without hand-crafted biological features)
StAC	Yes	No	Standard AttCRISPR

Note: The method with superscript of * and # is reported by Wang *et al.* and Muhammad Rafid *et al.* respectively. Specially, CRISPRpred(SEQ) take another set of hand-crafted sequence-based feature to improve performance.

a single experiment. The experiment is repeated ten times with the results recorded and averaged finally.

2.5 Experiment design

Two different experiments are carried out in our work. The first one is designed for ablation analysis of AttCRISPR. We compare the performance of end2end method (without any hand-crafted biological features) in both spatial and temporal domain. Furthermore, we test the ensemble method based on the same strategy to prove that the ensemble method in both spatial and temporal domain can significantly improve the performance.

The second experiment is designed to compare the performance of AttCRISPR with other current prediction methods. In order to make the comparison apples to apples, we reduce the dimensionality of the same hand-crafted biological features as DeepHF's, which has been shown to enhance the predictability of a deep-learning model greatly, with a multilayer perceptron. Then follow the Equation (14) to achieve the final prediction value. AttCRISPR (with the hand-crafted biological features) performs better on all three datasets than DeepHF.

Our baselines have a comprehensive coverage of the methods tested in these datasets. In Table 1, we annotate some properties of these baselines (is/isn't neural models, is/isn't end2end models) All of the experiments were carried out in Python 3.6 using Keras 2.2.4 and one GeForce RTX 2080Ti Super was used for training and testing if needed.

3 Results

We design experiments to address the following questions:

- In the absence of hand-crafted biological features, whether the stacking of method in spatial domain and temporal domain can get better performance then use these method alone? → Section 3.1
- How does AttCRISPR perform compared to current state-of-the-art methods, covering both conventional machine learning and deep-learning models? → Section 3.2
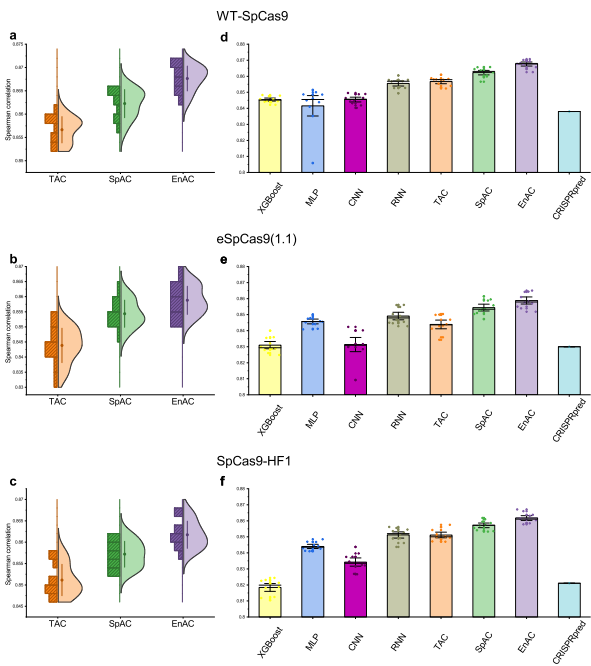

Fig. 6. In the absence of hand-crafted biological features, performance of different algorithms for sgRNA activity prediction. (a)-(c) The performance of Temporal AttCRISPR, Spatial AttCRISPR and Ensemble AttCRISPR. The half-violin plots show the mean and distribution of the Spearman correlation coefficient between predicted and measured sgRNA activity scores over all tests. (d)-(f) In the absence of hand-crafted biological features, the performance of all prediction methods in these datasets as far as we know. The $mean \pm s.d.$ of the Spearman correlation coefficient between predicted and measured sgRNA activity scores are shown in the bar plots.
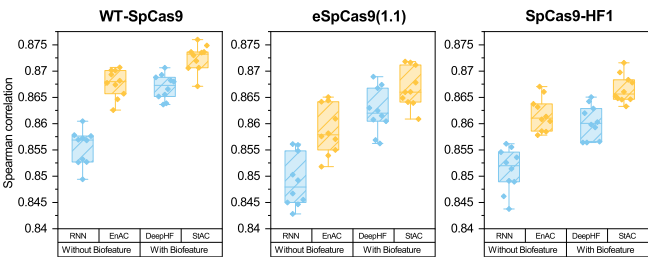

Fig. 7. Performance comparisons for the methods before and after integrating with hand-crafted biological features, where DeepHF is the RNN integrated with hand-crafted biological features, and StAC is the EnAC integrated with hand-crafted biological features. The box plot show the mean and distribution of Spearman correlation coefficient between predicted and measured sgRNA activity scores over all tests.

- How can researcher understand the decisions made by AttCRISPR locally and globally, based on attention mechanisms? → Section 3.3

3.1 Model building and stacking

In Table 2, we list the performance of methods in spatial or temporal domain and the stacking of methods. Temporal AttCRISPR, TAC for short, achieved Spearman correlation coefficients of 0.857, 0.844, 0.851 respectively in above three dataset. Spatial AttCRISPR, SpAC for short, corresponds to 0.862, 0.854, 0.857. In the absence of hand-crafted biological features. Ensemble AttCRISPR achieve the best performance of our knowledge, corresponds to 0.868, 0.859, 0.862.

sgRNA on-target activity prediction   **7**

Table 2. Performance comparisons for different methods in the absence of hand-crafted biological features (take Spearman correlation coefficient as evaluation index)

Method	WT-SpCas9	eSpCas9(1.1)	SpCas9-HF1
XGBoost*	0.845	0.831	0.818
MLP*	0.842	0.846	0.844
CNN*	0.846	0.831	0.834
RNN*	0.856	0.849	0.851
TAC	0.857	0.844	0.851
SpAC	0.862	0.854	0.857
EnAC	**0.868**	**0.859**	**0.862**
CRISPRpred(SEQ)#	0.838	0.830	0.821

Table 3. Performance comparisons for the methods before and after integrating with hand-crafted biological features (take Spearman correlation coefficient as evaluation index)

Method	WT-SpCas9		eSpCas9(1.1)		SpCas9-HF	
	Mean	Std Dev $(\times 10^{-3})$	Mean	Std Dev $(\times 10^{-3})$	Mean	Std Dev $(\times 10^{-3})$
RNN*	0.856	3.33	0.849	5.00	0.851	4.11
EnAC	0.868	2.66	0.859	4.66	0.862	3.19
DeepHF*	0.867	**2.37**	0.862	4.24	0.860	3.21
StAC	**0.872**	2.55	**0.867**	**3.71**	**0.867**	**2.65**

Note: The method with superscript of * and # is reported by Wang *et al.* and Muhammad Rafid *et al.* respectively. In the tables, we use the results reported in the relevant papers as the performance of the method directly.

In addition, in Table 2, the performance of other methods without using hand-crafted biological features, are also recorded. Regardless of the method we developed, RNN reported by Wang *et al.*, which can be categorized as the method in temporal domain, is the most predictive with Spearman correlation coefficients of 0.856, 0.849, 0.851. It's obvious that the ensemble AttCRISPR is better at prediction (Figure 6 (a-c)). Furthermore, the prediction ability of models could be boosted by addition of other hand-crafted biological features, which can't be obtained directly by sequence information.

Further experiment is designed to compare the performance of standard AttCRISPR (hand-crafted biological features are used to improve performance of ensemble AttCRISPR) and DeepHF, which is a current state-of-the-art method.

3.2 Performance comparison

In order to validate the conclusion that integrating with hand-crafted biological features can improve the predictive performance of methods, we follow Equation (14) to modify the ensemble method and design the control experiment using the same strategy. What's more, we compare the standard AttCRISPR and DeepHF (Table 3).

As shown in Table 3, in the absence of hand-crafted biological features, AttCRISPR has significant advantages over DeepHF in predictability. Further, integration with the hand-crafted biological features can also improve the performance of AttCRISPR, and achieve Spearman correlation coefficients of 0.872, 0.867 and 0.867 for WT-SpCas9, eSpCas9(1.1) and SpCas9-HF1, respectively. Meanwhile, DeepHF achieve 0.867, 0.862 and 0.860, respectively. After integrating with biological features, the performance gap between AttCRISPR and DeepHF is shortened, while AttCRISPR still has better performance. In addition, we also compare the standard deviation of data obtained in ten tests, which

are also shown in Table 3. It reveal that AttCRISPR is more stable than DeepHF.

3.3 Interpretability of the AttCRISPR

In the following sections, we will analyze the insight into activity of sgRNA bring through the attention mechanism at both global and local levels to validate the attention module in the AttCRISPR can help us to understand the decisions.

3.3.1 Global interpretability

At global level, an important question we expect AttCRISPR to answer is which nucleotide it prefers at each position on the sequence. In fact, Wang *et al.* has already answered this question in detail with the DeepSHAP method. Leaving aside nucleotides, just consider the importance of each position, the method of Liu *et al.* have gone some way too. While our method is not based on the post hoc explanations techniques and input perturbations, the only work we need to do is to get the first-order preference A generated by the attention module. Specifically, we use the first-order preference A_s generated by the spatial attention module instead of $\tilde{A}$ generated by the temporal attention module. The latter is in a higher dimensional dense space which make it difficult to understand. In practice, we input every sgRNA into the spatial AttCRISPR, in order to obtain the A_s from the spatial attention module and take its mean value. Then we rescale it through Z-score to obtain a standardization value and the final result is shown in Figure 8.

As shown in Figure 8, we captured the preference for each position-dependent nucleotide on the sgRNA sequence. The result revealed that A and G typically have a positive contribution to the activity of sgRNA, while T typically have a negative contribution. It is agree with previous conlusion that when Cas9 is binding sgRNA, it prefer the one containing purines to pyrimidines (Wang *et al.*, 2014). In addition, global interpretability also pointed out that distinct from other nucleotides, G is strongly favored at position 20. This is consistent with the conclusions of several other reports (Wong *et al.*, 2015; Doench *et al.*, 2014).

Furthermore, the preference of the nucleotide at the same position doesn't change dramatically with the Cas9 nucleases, while we still notice that compared with the other two datasets, C makes a more positive contribution to the activity of sgRNA with the SpCas9-HF1 especially in the position 5, which is evident in Figure 8 (d-f).

The above discussion shows that, in the task of sgRNA activity prediction, attention mechanism can help us understand the decision made by AttCRISPR and reveal the insight into activity of sgRNA.

3.3.2 Local interpretability

At local level, we analyze a case (consisting of three sgRNAs as Table 4 show), then we expect AttCRISPR to answer two imortant questions based on the local interpretability. First, how can we optimize a sgRNA to have more on-target activity. Second, what are the reasons for the low activity of the sgRNA.

For the first question, we input the least-active sgRNA in Table 4 (with the index of 8493 and the activity of 0.831, call source sgRNA for convenience) into AttCRISPR. The score of each position is obtained based on Equation (6) (the calculated symbol with W is Hadamard product instead of dot product, in order to achieve the result in vector form), and the results are shown in Figure 9. Acording to Figure 9, scores at position 14 and 16 of source sgRNA are significantly below the base line (in fact, the scores at position 6 and 11 is also noteworthy, however, we don't find sgRNA in the dataset for comparison). If replaced the T at position 14 with C, would generate the same sgRNA as the one with index of 8491, which is with a activity value of 0.861. If replaced the T at position 16 with C, would generate the same sgRNA as the one
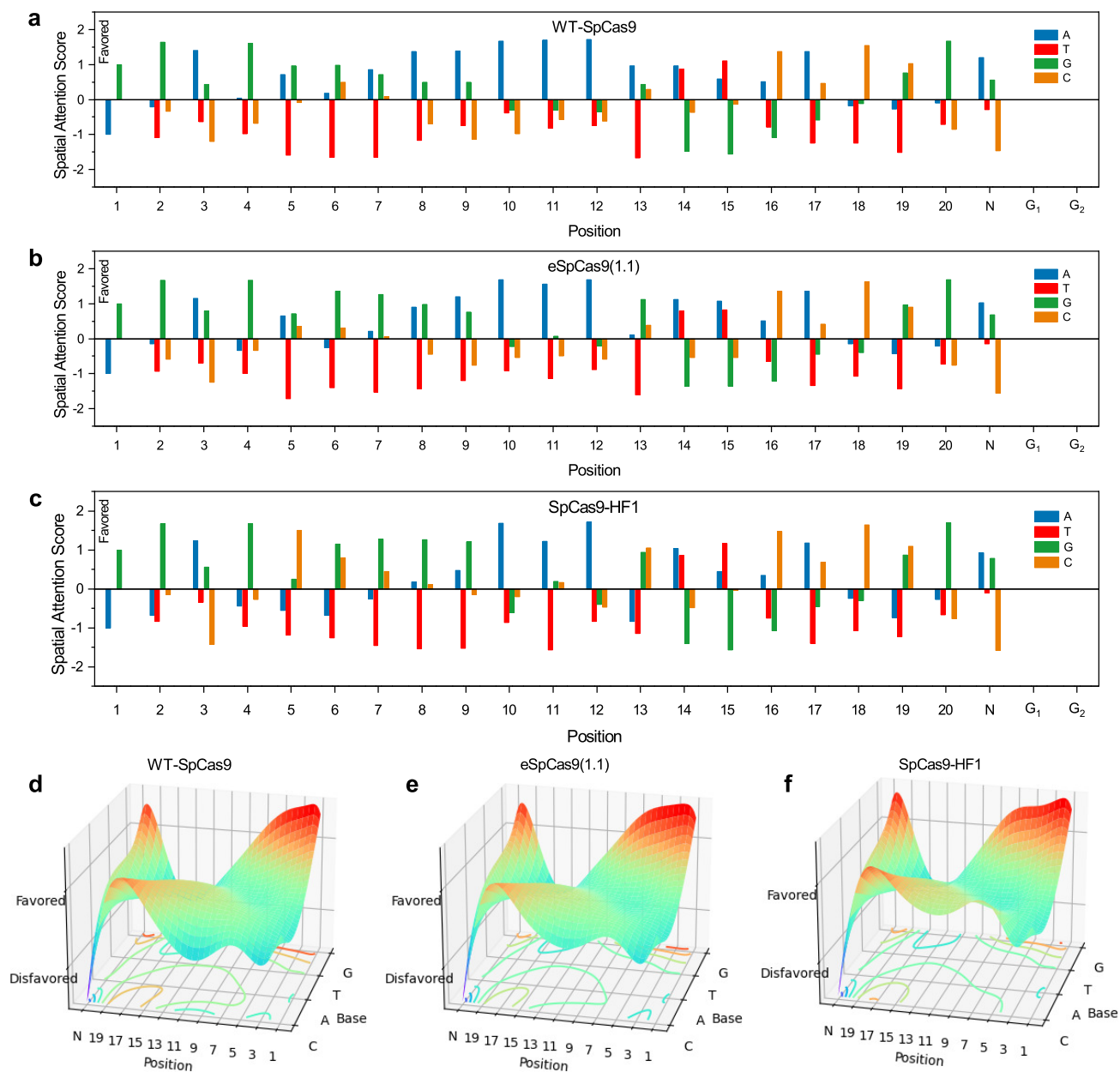
Fig. 8. Preference for each position-dependent nucleotide on the sgRNA sequence. (a)-(c) Bars show the score of preference after standardization, and the higher the number, the more positive it is for activity of sgRNA. The numbers below indicated the position of the nucleotides on-target DNA. (d)-(f) Preference surfaces for each position-dependent nucleotide fitted with Bézier surfaces. Each position-dependent nucleotide is a control point. The coordinates of the control points on the vertical axis represent the degree of preference, and the higher the position-dependent nucleotide corresponding to the control point is, the more positive it is for the activity of sgRNA. The contour plots at the bottom show the area of position-dependent nucleotides which have different contribution to the activity of sgRNA.

with index of 8492, which is with a activity value of 0.869. Therefore, we could conclude that the local interpretability is helpful for us to optimize the sgRNA without exhaustively search.

The second question we expect AttCRISPR to answer is why it give two low scores at position 14 and 16. In practice, we will try to answer this question with second-order preference. Let AttCRISPR output the second-order preference matrix B corresponding to the source sgRNA, and show it in Figure 10. It is obvious that a few unusual bright spots appear in the red box in Figure 10, which show that the nucleotide at position 15 has a great effect on the score of position 14 and 16 (instead of position 13 or

17, which corresponding position are relatively dim in color). As shown in Table 4, in source sgRNA there are three consecutive Ts at position 14, 15 and 16, and this may reveal that multiple consecutive Us on sgRNA would leads to low on-target activity of sgRNA, which is consistent with earlier report (Wu *et al.*, 2014).

4 Discussion

In this article, we have developed a new prediction method, called AttCRISPR for the activity of sgRNA. We take the ensemble of both spatial

Table 4. Three sgRNA and their activity

Index	sgRNA	Activity
8493	ACATGACTTTGGATTTCCCCAGG	0.831
8492	ACATGACTTTGGATTCCCCCAGG	0.869
8491	ACATGACTTTGGACTTCCCCAGG	0.861

Note: *Activity* of sgRNA represents the activity reported in WT-SpCas9 dataset (Supplementary Data 1).
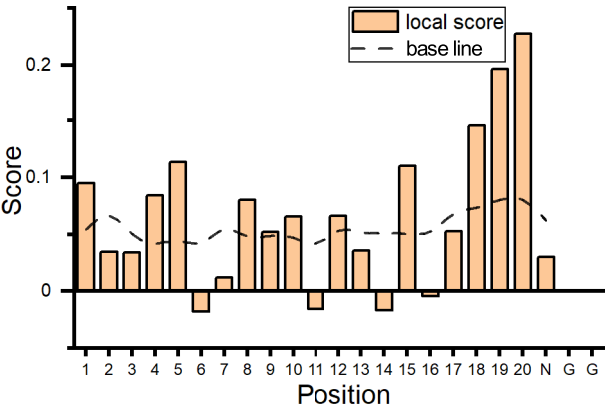

Fig. 9. The scores of sgRNA with index of 8493 obtained by the temporal AttCRISPR. The bar plot reveals the score at each position at local level, while the line of dashes reveals the scores at global level. At global level we calculated the scores of all sgRNA, then average the scores at each position to obtain the global level which can be used as base line and is helpful to judge the relative height of the scores.

Fig. 10. The visualization of second-order preference matrix B, the elements in the i-th row and the j-th column represent the influence of nucleotide at position j when generate the first-order preference $\tilde{A}$ at position i. The warmer the color, the more important it is. In the red box, a few unusual bright spots appear. To be more specific, the nucleotide at position 15 has a great effect on the first-order preference at positon 14 and 16.

and temporal domain to predict the on-target activity of sgRNA. Throught ablation analysis and testing a series of possible network structure, we demonstrate that the ensemble method performs better than other methods on this task. In addition, we apply attention modules in both the spatial and temporal parts of AttCRISPR, and design two experiments combined with some early reports to prove that attention mechanisms can help researcher understand the decisions made by model which make it easy to optimize low activity sgRNA without exhaustively search.

As shown in Figure 10, we note that the brightness at coordinates (14, 15) and (16, 15) exceeds (14, 13) and (16, 17). In other words, there are two unusual bright spots at coordinates (14, 15) and (16, 15). This could explain that the nucleotide trimer at position 14, 15, 16 has a great influence on the decision made by AttCRISPR. We believe that we can use a carefully designed 3×3 convolution kernel, and move it along the diagonal of the second-order preference matrix B, in order to find all kind of nucleotide trimer that have a great influence on the decision made by AttCRISPR. Further experiments may be need to validate.

Current architecture of AttCRISPR focuses on predicting the on-target activity of conventional sgRNA which have a PAM based on NGG. However it can be extended to other Cas9 species, variants or off-target task easily.

5 Conclusion

In this paper, we develop AttCRISPR, an ensemble method of both spatial method and temporal method follow the stacking strategy with strong interpretability. AttCRISPR prove that the ensemble methods have a better performance in the datasets of DeepHF and can compete with current state-of-the-art methods. In addition, AttCRISPR apply attention mechanisms in 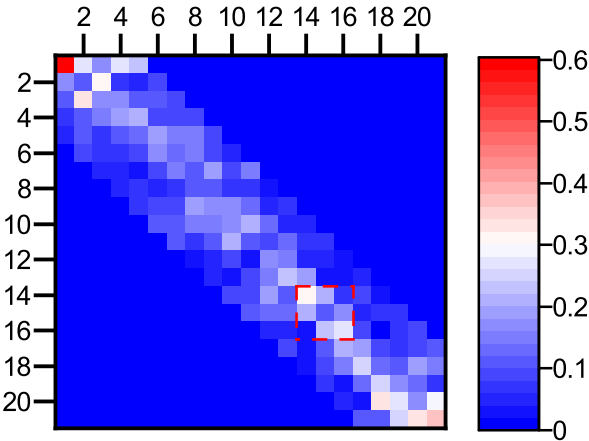both the temporal and spatial part, and we explain the decisions made by AttCRISPR through the attention module which is consistent with earlier reports. Further, we also discovered that the output of the attention module can be used to optimize the low-activity sgRNA without exhaustively search, and our optimization results are verified with available experimental data. To our knowledge, this is the first time that the attentional mechanim is used to reveal the reasons influencing on-target activity of sgRNA, and assist sgRNA design.

Conflict of Interest: none declared.

References

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**(1), 1–18.

Cong, L., Ran, F. A., Cox, D. D., Lin, S., Barretto, R. P. J., Habib, N., Hsu, P., Wu, X., Jiang, W., Marraffini, L. A., *et al.* (2013). Multiplex genome engineering using crispr/cas systems. *Science*, **339**(6121), 819–823.

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**(12), 1262–1267.

Ishii, T. (2017). Reproductive medicine involving genome editing: clinical uncertainties and embryological needs. *Reprod. Biomed. Online*, **34**(1), 27–31.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, **337**(6096), 816–821.

Kang, X., He, W., Huang, Y., Yu, Q., Chen, Y., Gao, X., Sun, X., and Fan, Y. (2016). Introducing precise genetic modifications into human 3pn embryos by crispr/cas-mediated genome editing. *J. Assist. Reprod. Genet.*, **33**(5), 581–588.

Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., Lee, S., Yoon, S., and Kim, H. (2018). Deep learning improves prediction of crispr-cpf1 guide rna activity. *Nat. Biotechnol.*, **36**(3), 239–241.

Kim, H. K., Kim, Y., Lee, S., Min, S., Bae, J. Y., Choi, J. W., Park, J., Jung, D., Yoon, S., and Kim, H. H. (2019). Spcas9 activity prediction by deepspcas9, a deep learning–based model with high generalization performance. *Sci. Adv.*, **5**(11), eaax9249.

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., and Joung, J. K. (2016). High-fidelity crispr–cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**(7587), 490–495.

Liang, P., Xu, Y., Zhang, X., Ding, C., Huang, R., Zhang, Z., Lv, J., Xie, X., Chen, Y., Li, Y., *et al.* (2015). Crispr/cas9-mediated gene editing in human tripronuclear zygotes. *Protein & Cell*, **6**(5), 363–372.

Lin, J. and Wong, K.-C. (2018). Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics*, **34**(17), i656–i663.

Liu, G., Zhang, Y., and Zhang, T. (2019a). Computational approaches for effective crispr guide rna design and evaluation. *Comput. Struct. Biotechnol. J.*, **18**, 35–44.

Liu, Q., He, D., and Xie, L. (2019b). 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**(10), e1007480–e1007480.

Liu, Q., Cheng, X., Liu, G., Li, B., and Liu, X. (2020). Deep learning improves the ability of sgrna off-target propensity prediction. *BMC Bioinformatics*, **21**(1), 51.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *arXiv e-prints*, page arXiv:1508.04025.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., Norville, J. E., and Church, G. M. (2013). RNA-guided human genome engineering via cas9. *Science*, **339**(6121), 823–826.

Muhammad Rafid, A. H., Toufikuzzaman, M., Rahman, M. S., and Rahman, M. S. (2020). Crisprpred(seq): a sequence-based method for sgrna on target activity prediction using traditional machine learning. *BMC Bioinformatics*, **21**(1), 223.

Rubeis, G. and Steger, F. (2018). Risks and benefits of human germline genome editing: An ethical analysis. *Asian Bioethics Review*, **10**(2), 133–141.

Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X., and Zhang, F. (2016). Rationally engineered cas9 nucleases with improved specificity. *Science*,

351(6268), 84–88.

Song, M., Kim, H. K., Lee, S., Kim, Y., Seo, S.-Y., Park, J., Choi, J. W., Jang, H., Shin, J. H., Min, S., Quan, Z., Kim, J. H., Kang, H. C., Yoon, S., and Kim, H. H. (2020). Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nat. Biotechnol.*, **38**(9), 1037–1043.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv e-prints*, page arXiv:1706.03762.

Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., *et al.* (2019). Optimized crispr guide rna design for two high-fidelity cas9 variants by deep learning. *Nat. Communications*, **10**(1), 4284–4284.

Wang, S., Peng, J., Ma, J., and Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.*, **6**(1), 18962.

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

Wong, N., Liu, W., and Wang, X. (2015). Wu-crispr: characteristics of functional guide rnas for the crispr/cas9 system. *Genome Biology*, **16**(1), 218.

Woo, S., Park, J., Lee, J., and Kweon, I. S. (2018). CBAM: convolutional block attention module. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Proceedings of the European Conference on Computer Vision (ECCV) 2018*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Wu, X., Scott, D. A., Kriz, A. J., Chiu, A. C., Hsu, P., Dadon, D. B., Cheng, A. W., Trevino, A. E., Konermann, S., Chen, S., *et al.* (2014). Genome-wide binding of the crispr endonuclease cas9 in mammalian cells. *Nat. Biotechnol.*, **32**(7), 670–676.

Zhang, G., Dai, Z., and Dai, X. (2020). C-rnncrispr: Prediction of crispr/cas9 sgrna activity using convolutional and recurrent neural networks. *Comput. Struct. Biotechnol. J.*, **18**, 344–354.