Gene 910 (2024) 148330

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

BERT-siRNA: siRNA target prediction based on BERT pre-trained interpretable model

Jiayu Xu [a], Nan Xu [b,c], Weixin Xie [a], Chengkui Zhao [a,c,*], Lei Yu [b,c,*], Weixing Feng [a,*]

[a] Institute of Intelligent System and Bioinformatics, College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
[b] Institute of Biomedical Engineering and Technology, Shanghai Engineering Research Center of Molecular Therapeutics and New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, No, 3663 North Zhongshan Road, Shanghai 200065, China
[c] Shanghai Unicar-Therapy Bio-medicine Technology Co., Ltd, No 1525 Minqiang Road, Shanghai 201612, China

ARTICLE INFO

Edited by: Stephen Kwok-Wing Tsui

Keywords:
siRNA prediction
BERT
Explainable deep learning
SARS-CoV-2

ABSTRACT

Silencing mRNA through siRNA is vital for RNA interference (RNAi), necessitating accurate computational methods for siRNA selection. Current approaches, relying on machine learning, often face challenges with large data requirements and intricate data preprocessing, leading to reduced accuracy. To address this challenge, we propose a BERT model-based siRNA target gene knockdown efficiency prediction method called BERT-siRNA, which consists of a pre-trained DNA-BERT module and Multilayer Perceptron module. It applies the concept of transfer learning to avoid the limitation of a small sample size and the need for extensive preprocessing processes. By fine-tuning on various siRNA datasets after pretraining on extensive genomic data using DNA-BERT to enhance predictive capabilities. Our model clearly outperforms all existing siRNA prediction models through testing on the independent public siRNA dataset. Furthermore, the model's consistent predictions of high-efficiency siRNA knockdown for SARS-CoV-2, as well as its alignment with experimental results for *PDCD1*, *CD38*, and *IL6*, demonstrate the reliability and stability of the model. In addition, the attention scores for all 19-nt positions in the dataset indicate that the model's attention is predominantly focused on the 5′ end of the siRNA. The step-by-step visualization of the hidden layer's classification progressively clarified and explained the effective feature extraction of the MLP layer. The explainability of model by analysis the attention scores and hidden layers is also our main purpose in this work, making it more explainable and reliable for biological researchers.

1. Introduction

Knockdown refers to weakening or reducing the expression level of target genes through various technical means. In 1984, RNA interference (RNAi) was born with the confirmation that introduction of a fragment of RNA into cells could change the expression of endogenous genes (Izant & Weintraub, 1984). In 1998, Andrew Fire and other researchers reported that the interference effect of double-stranded RNA was stronger than that of single-stranded RNA (Fire et al., 1998). Small interfering RNA (siRNA) can be used as molecular tools to inhibit gene expression and design new drugs. By introducing appropriate siRNA molecules, siRNA specifically binds to the mRNA of the target gene, leading to the degradation of the target mRNA, thereby inhibiting the expression of the target gene and achieving knockdown of the gene. The first siRNA drug (Onpattro) was approved in 2018 by the Food and Drug Administration (FDA) for the treatment of polyneuropathy in patients

Abbreviations: RNA, Ribonucleic Acid; mRNA, Messenger RNA; siRNA, Small interfering RNA; RNAi, RNA interference; DNA, Deoxyribonucleic Acid; SARS-CoV-2, Severe Acute Respiratory Syndrome – 2 COVID – 19; PDCD1 (CD279/PD-1), Programmed Cell Death 1; CTLA-4, Cytotoxic T-Lymphocyte Antigen-4; CD38, Cluster of Differentiation 38; IL6, Interleukin 6; FDA, The Food and Drug Administration; MLP, Multilayer Perceptron; MSE, The mean square error; AUROC, Area Under the Receiver Operating Characteristics Curve; AUPRC, Area Under the Precision Recall Curve.

* Corresponding authors at: Institute of Intelligent System and Bioinformatics, College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China (C. Zhao, W. Feng). Institute of Biomedical Engineering and Technology, Shanghai Engineering Research Center of Molecular Therapeutics and New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, No, 3663 North Zhongshan Road, Shanghai 200065, China (L. Yu).

E-mail addresses: xujiayu@hrbeu.edu.cn (J. Xu), xunanlouis@126.com (N. Xu), xieweixin@hrbeu.edu.cn (W. Xie), zhaochengkui@hrbeu.edu.cn (C. Zhao), yulei@nbic.ecnu.edu.cn (L. Yu), fengweixing@hrbeu.edu.cn (W. Feng).

https://doi.org/10.1016/j.gene.2024.148330
Received 12 December 2023; Received in revised form 22 February 2024; Accepted 28 February 2024
Available online 29 February 2024

with hereditary transthyretin-mediated amyloidosis (Rossi & Rossi, 2021). Also, siRNA showed immunomodulatory potential in cancer treatment by down-regulating immune-suppressive proteins. For example, *PD-1* and *CTLA-4* restrict immune cell function and pose challenges to cancer immunotherapy (Monty et al., 2021). According to recent research, it could also be used to against SARS-CoV-2 by targeting the highly conserved region of SARS-CoV-2 to inhibit a wide spectrum of viral variants (Chang et al., 2022).

Comparing with our previous research on microRNA-based shRNA (C. Zhao et al., 2022; C Zhao et al., 2022), the siRNA molecule is shorter, easier to synthetize and more directly transduced, which makes siRNA technology an efficient and safe alternative for drug development.

Since the main factor that determines efficiency in siRNA knockdown is the binding site in mRNA, identifying such targets is of prime importance. However, experimental screening of thousands of nucleotides is costly. Alternatively, selection of optimal targets in mRNA can be performed using statistical and machine learning methods that screen all possible binding sites and predict knockdown siRNA efficiency. These methods greatly optimize resources for further experimental evaluation.

Machine learning methods have been used for this purpose using increasingly available experimentally validated siRNA datasets. Sætrom first collected all independent experimental siRNA datasets to classify the high and low efficient siRNA with a binary classification model based on Genetic Programming (GP) algorithm (Saetrom, 2004). Reynolds systematically analyzed 180 siRNAs targeting two mRNAs (Reynolds et al., 2004), whereas Huesken worked on 2431 siRNAs datasets using high-throughput technology, thus expanding the scale of experimental data by about ten-fold (Huesken et al., 2005). They also developed a neural network algorithm 'Biopredsi' with the new dataset. Shabalina added thermodynamic features to the input of the machine learning model, improving efficiency in prediction (Shabalina, Spiridonov, & Ogurtsov, 2006). Vert and Ichihara established two simple linear models, DSIR and i-Score, to predict siRNA knockdown efficiency (Ichihara et al., 2007; Vert, Foveau, Lajaunie, & Vandenbrouck, 2006). These existing models require additional relevant features besides the raw sequence features, making the preprocessing process complex. By adopting deep learning, we circumvent this cumbersome preprocessing, simplifying the approach.

Traditional machine learning algorithms can quantitatively predict knockdown efficiency using neural networks, support vector machine or random forest, which extract biological features from siRNA. Despite all these efforts, the relatively low prediction accuracy cannot satisfy our experiment design. Compared with the hundreds of billions ($4e^{19}$) of possible siRNAs, the training data is only of a few thousands of siRNAs. Thus, we propose that the prediction model could be further improved and efficiently make full use of the limited training data with the help of transfer knowledge from the billions of human genomic sequences.

Advances in natural languages have enabled training of complex models understanding the semantics from unlabeled text. The BERT model is one excellent model using this technique (Devlin, Chang, Lee, & Toutanova, 2018). It is trained to predict words masked out in a sentence or to predict the next word or sentence following the previous context. Applied to the genomic sequences, the DNABERT model is developed to capture global and transferable understanding of genome sequences based on up- and down-stream nucleotide contexts (Ji, Zhou, Liu, & Davuluri, 2021). This model has shown to be useful to predict promoters, splice sites and transcription factor binding sites. Therefore, we hypothesize that a pretrained DNA sequence model could be used for siRNA prediction, since both DNA and RNA store genetic information and RNA is transcribed from DNA.

In addition, the ability to interpret what a model has learned is receiving an increasing amount of attention (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019; Reynolds et al., 2004). Investigating the explainability of the deep learning model is also our main purpose in this work. It is crucial for human to understand how the deep learning model functions and whether we could trust the prediction based on the knowledge we learn in the model inner parameter. The model could be trusted if the decision is made based on the factor or region that is in correspondence with the intuition of the biological researchers.

In this study, we propose an explainable BERT-based model for the siRNA target prediction. This model takes advantage of pre-training from a large number of human genomic sequences, leading to state-of-the-art performance in siRNA prediction. To show robustness and pivotal biological meaning of this model, we verified it on potent siRNA against SARS-CoV-2 and knockdown experimental data of immune-related genes *PDCD1*, *CD38*, and *IL6*. The knockdown experiments for the *PDCD1*, *CD38* and *IL6* were conducted in our laboratory. Finally, we investigate the explainability of the model by extracting attention score and intermediate hidden layers, making it more explainable and reliable for biological researchers.

2. Material and methods

2.1. Public siRNA target knockdown data

The final training dataset was the largest siRNA dataset from Huesken (Huesken et al., 2005), consisting of 2431 experimental siRNA knockdown samples for 34 mRNA sequences. This dataset was divided into training, validation, and test dataset at a ratio of 8:1:1. The validation dataset was used to optimize hyperparameters and the test dataset serves as an internal test dataset for evaluating model performance.

The other two largest siRNA datasets from Reynold (Reynolds et al., 2004) and Katoh (Katoh & Suzuki, 2007), which included 914 samples, were combined together as the independent test dataset. Reynold dataset contains 252 samples and Katoh dataset contains 662 samples. For each sample, it includes 19-nt siRNA sequence with efficiency value ranging from 0 to 1.

2.2. Experimental verification dataset

Coronavirus has the largest genome among known RNA viruses. In order to identify highly efficient and specific siRNA sequences against SARS-CoV-2 variants, Chang (Chang et al., 2022) conducted a comprehensive screening of siRNA sequences. The selected siRNA sequences exhibit a high coverage rate across the SARS-CoV-2 genome, targeting various regions within the genome, all while demonstrating low propensity for secondary structure formation in their respective targeting regions. Furthermore, these siRNAs excluded those with a high potential for biased effects on the human transcriptome and targeting genes essential for cell survival. Ultimately, the top 11 siRNAs with the lowest predicted off-target effects and highest predicted efficacy were selected.

Considering the excellence and reliability of Chang's study on siRNA knockdown in SARS-CoV-2, we compared the knockdown results of the 11 siRNAs from Chang's study with our model predictions to verify the reliability and accuracy of our model.

In addition to verifying the model's performance using experimental results from existing studies, we also utilized our own laboratory data to verify the accuracy of the model predictions.

PDCD1, *CD38*, and *IL6* are crucial participants in immune regulation and are among the genes widely studied in current medical research. For instance, PD-1 inhibitors, IL-6 inhibitors, and CD38 inhibitors have been widely applied in various cancer types. Therefore, we collected the three genes *PDCD1*, *CD38* and *IL6* as representative genes for knockdown experiments with 8, 5 and 5 siRNAs respectively to independently verify the performance of our model.

2.3. Bert-based model for the siRNA prediction

Based on the Bidirectional Encoder Representations from Transformers (BERT) model structure and DNABERT model (Devlin et al., 2018; Ji et al., 2021), we developed the BERT-siRNA, which was

pretrained with the up- and down-stream nucleotide contexts of the human genome to capture global and transferrable understanding of genomic sequences.

The BERT-siRNA model's structure is shown in Fig. 1. BERT-siRNA takes a set of the siRNA sequences and transforms them into k-mer tokens. Our model uses k as 6, which means that each group of 6 bases in the siRNA sequence sequentially forms a token required by the model. For example, if the original sequence is ATCGAATCGAACCGAACA, the first token is ATCGAA, the second is TCGAAC, and the third token is CGAACA. This approach is employed to capture local information within the siRNA sequence. Compared to representing individual nucleotides, this method provides a more comprehensive and focused information, contributing to the enhancement of model performance. The special tokens 'CLS' and 'SEP' are in the start and end separately in the input. The input is then embedded into a numeric vector. The pre-trained BERT model contains 12 attention heads, 768 hidden units, and 12 transformer encoder blocks. This model structure mainly features for the attention head. The query (q), key (k), and value (v) vector are first calculated from each input hidden state h. The attention head output z at position i is calculated as follows:

$$z^i = softmax\left(\frac{q^{(i)} \bullet K}{\sqrt{d_{head}}}\right) \bullet V \tag{1}$$

where d_{head} is the dimension of the key vectors, K is the key matrix, and the V is the value matrix. 12 such attention heads are used in one encoder layer. The outputs of attention heads are concatenated to get the final output vector with 768 dimensions for each nucleotide. It should be denoted that the attention scores representing the relevance of information between two positions are calculated by the multiplication of the q and k vectors. The attention scores could be used to explain the model prediction on the specific task.

In our model structure (Fig. 1), we extract 768-dimension vector from the last position, 'CLS' output, from the pretrained BERT model. Then the vector is processed by the Multilayer Perceptron (MLP) block including one dropout layer, three dense layers and one sigmoid output layer. All the siRNA sequences contain 19 nucleotides, and are processed by a 6-mer window to get 16 tokens with the 'CLS' and 'SEP'. The final output is a sigmoid output ranging from 0 to 1 representing predicted knockdown efficiency.

Hyperparameters include dropout rate, number of dense layers, learning rate, and the weight decay. They were determined through training and validation with the Huesken dataset. Finally, we used all the best hyperparameters to train the final model with the Huesken dataset.

2.4. Training and evaluation metrics

We used the Huesken training dataset and validation dataset to find the best hyperparameters. Then the Huesken test dataset was used to evaluate the model performance. To compare with all the other methods, we finally trained our model on the Huesken dataset with the selected hyperparameters. The Reynold and Katoh dataset were combined together as the independent test dataset for all the algorithms when comparing their performance.

The mean square error (MSE) between siRNA knockdown efficiency and model output was set as the loss to evaluate the model during training process. MSE is an indicator that measures the difference between predicted values and actual values. For predicted values $\widetilde{y}_i$ and actual values y_i of the i-th sample, the MSE calculation formula is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}_i - y_i)^2 \tag{2}$$

where n is the number of samples. The smaller the MSE, the better the model fits the sample and the smaller the difference. The Pearson's and Spearman's correlations between the knockdown efficiency and model's predictions were both used as the evaluation metrics for the regression model. Using correlations between predicted and true values can provide information about model performance and prediction accuracy. If the correlation is high, it means that the model's predicted value and the true value have a linear relationship to a certain extent, indicating that the model captures the overall trend better.

After each training iteration, validation sets were used to evaluate the performance of the model. Different hyperparameters have different impacts on the model. By monitoring the performance on the validation set, we could adjust the model hyperparameters in time, including the learning rate, regularization (dropout rate and weight decay) and number of dense layers.

The learning rate determines the response of the model to gradient descent during the training process and determines the convergence effect of the model. The regularization helps improve the generalization ability of the model and reduce the risk of overfitting. Dropout rate is a regularization technique during neural network training. In each training iteration, some neurons are randomly selected and their outputs are set to zero (that is, discarded) preventing the model from relying too

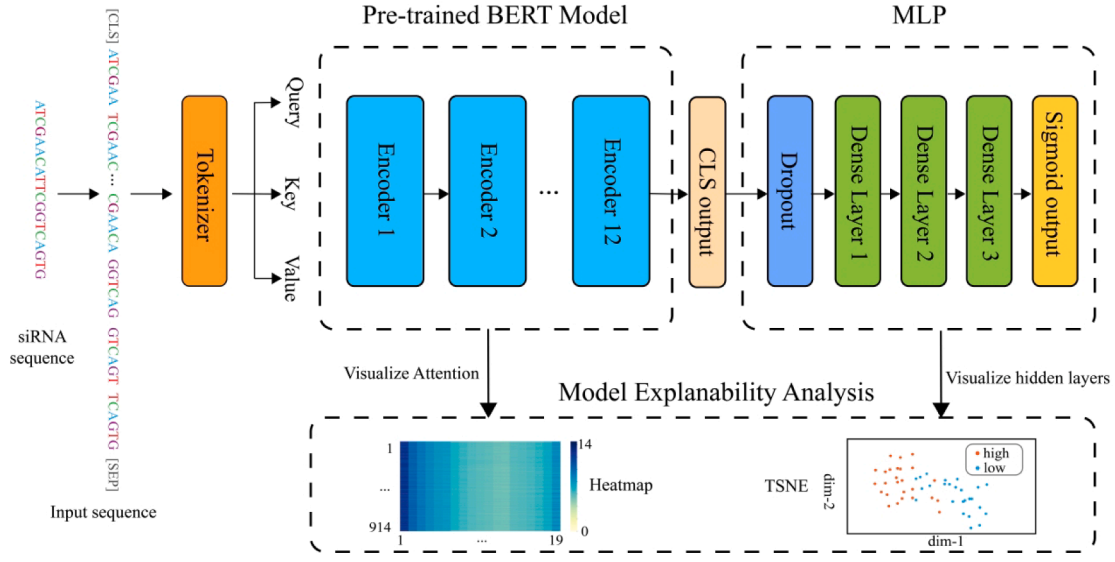

Fig. 1. Description of the BERT-siRNA architecture and the model explainability analysis.

heavily on certain input features, thereby enhancing the model's generalization ability. During the testing phase, all neurons are retained, but their outputs are scaled according to their probabilities during training. Weight Decay controls the complexity of the model by adding a weight penalty term to the loss function. It limits the complexity of the model by penalizing excessive weights, thereby preventing overfitting. Increasing the number of dense layers can enable the model to adapt to more complex nonlinear relationships, enabling it to better fit the data and learn more intricate patterns. However, adding more dense layers also escalates the model's parameter count and computational complexity, which may lead to overfitting issues. Therefore, we should search for an appropriate number of dense layers to meet the requirements of the task at hand.

We employ the method of controlling variables to adjust parameters and experiment with different parameter combinations. Through iterative training on the training dataset and evaluation on the internal validation dataset from Huesken, the set of parameters yielding the best evaluation results was ultimately chosen as the final model parameters, enabling the model to achieve peak performance on the validation dataset. Then the Huesken test dataset is used to evaluate model performance.

Then, we split the test dataset into high and low knockdown groups with efficiency greater than 0.7 or not. The Area Under the Receiver Operating Characteristics Curve (AUROC) and Area Under the Precision Recall Curve (AUPRC) are used as the evaluation metrics for the classification task.

While, due to the data imbalance between the high and low groups, which the amount of the high knockdown siRNAs is far less than the low, the AUPRC is more suitable for the classification evaluation with less impacted by the data imbalance. Finally, for the well-trained model applied for the classification task, the siRNA is predicted as high knockdown with output score greater than 0.7.

2.5. Explainability of the model

The model explainability is also important for the machine learning model besides the outstanding predictive power. To explore the mechanism for the deep learning model, we adopted two explainable ways. We first investigated the model's explainability by visualizing its attention scores. Models typically focus more strongly on features that are more important or informative for the task at hand while giving higher weight to specific parts of the input. The attention scores reflect the degree to which the model pays attention to different parts of the input when making predictions, making the model's prediction results more explainable. And visualization of attention scores allows us to intuitively understand the focus of the model. The nucleotide-level attention scores are calculated according to the DNABERT-viz. Formally, let q^* be the query vector of the 'CLS' token. Let k_j be the key vector for the j-th k-mer token, $j \in (1, \cdots, 14)$, where the siRNA sequence has 14 tokens. Then the attention score for each k-mer token over all the attention heads H is as follows:

$$\alpha_j = \sum_{h=1}^{H} \frac{exp\left(\frac{q^{*T} k_j}{\sqrt{d}}\right)}{\sum_{l=1}^{14} exp\left(\frac{q^{*T} k_l}{\sqrt{d}}\right)} \tag{3}$$

The attention score for each nucleotide is the average of all the k-mer attention scores α_j that contain this nucleotide. Secondly, the hidden states for the last layers could also give us insight into the model prediction. We output the final 3 hidden layers in the MLP to show the discriminative power across different layers. After unsupervised clustering, the distance between high and low knockdown groups are calculated and shown in the TSNE plot. The distance between groups could be used as the metric to evaluate the performance of the features extracted by the hidden layers.

2.6. Knock-down experiment

We used Jurkat cell line, which is an acute T cell lymphocytic leukemia cell. From this cell line, we randomly selected three genes that are widely studied in the biomedical field: *PDCD1*, *CD38*, and *IL6*. Knockdown experimental data for these genes were obtained from our laboratory. We have generated stable selected genes expression with subsequent transduction of selected genes containing lentivirus on Jurkat cell line for assessment of upregulation and downregulation of corresponding gene. Afterwards, the corresponding selected genes overexpression of Jurkat cells were detected by flow cytometry. Residual genes expression of stable Jurkat cells for selected genes have authenticated by BioLegend, San Diego, CA, USA.

3. Results

3.1. Model hyperparameters evaluation

To obtain the best performance of the deep learning model, we tuned four hyperparameters, initializing MLP layers dimension (2), learning rate ($1e^{-5}$), dropout rate (0.1) and weight decay rate (0). The minimum of the MSE loss for the validation dataset during training was used to evaluate the performance of the model with the hyperparameters shown in Fig. 2. Best performance was obtained with MLP layers of dimension 4 (Fig. 2A), indicating that a network with too deep or too shallow structure does not work well. Performance was increased with learning rate $5e^{-5}$ and dropout rate 0.1 (Fig. 2B, C). The weight decay adding L2 weights penalty to the cost function helped prevent overfitting the training data, and led to the best validation performance with $1e^{-2}$ (Fig. 2D). The selected hyperparameters were used to train the best deep learning model.

3.2. The influence of the pretraining

Better prediction on siRNA knockdown efficiency is achieved by model pretraining with the DNA sequences, including nucleotide preference and genomic structural information into the model. The model performance comparison, with or without pretraining using the validation dataset (Fig. 3A), shows that including pretraining led to a faster and better performance (Pearson correlation 0.57 vs. 0.49).

We used our selected hyperparameters and pretraining model to train the model on the Huesken training dataset, whereas the best model is selected by the validation dataset. As shown in Fig. 3B, the model prediction on the unused Huesken test dataset shows the model's predictive ability (Pearson correlation = 0.636).

3.3. Model performance comparison with existing algorithms

To benchmark the siRNA prediction models, we combined the two largest independent siRNA knockdown datasets, Reynold and Katoh, to represent our benchmark dataset. All models are trained based on the whole Huesken dataset. The Spearman and Pearson correlation is used as the metric for the regression model performance. The BERT-siRNA model outperformed DSIR - the best one of other models (Spearman's correlation 0.639 vs. 0.598, Pearson's correlation 0.611 vs. 0.579) (Fig. 4A). To compare the binary classification performance of the model, AUROC and AUPRC were used to evaluate the discriminative ability. Due to the imbalance of the data, the AUPRC and AUROC are more suitable as the main metrics to evaluate the model's classification performance. BERT-siRNA performed better than i-Score, the best of others (AUPRC 0.714 vs. 0.674, AUROC 0.826 vs. 0.804) (Fig. 4B, 4C).

Overall, the regression and classification model comparison show that our model could give a better prediction on the knockdown efficiency and has better ability to distinguish the potent siRNAs.

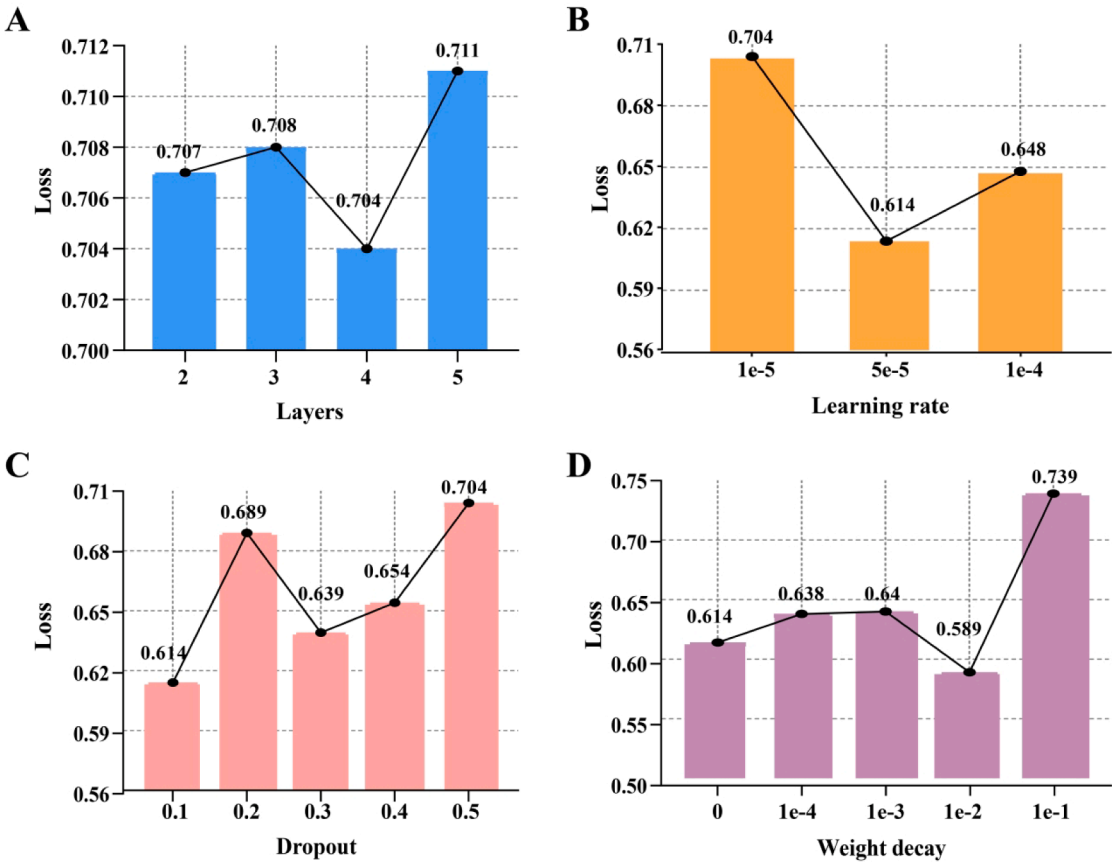

Fig. 2. Model performance with different hyperparameters: (A) number of the MLP layers; (B) learning rate; (C) dropout rate in the MLP dropout layer; (D) L2 regularization weight decay rate.

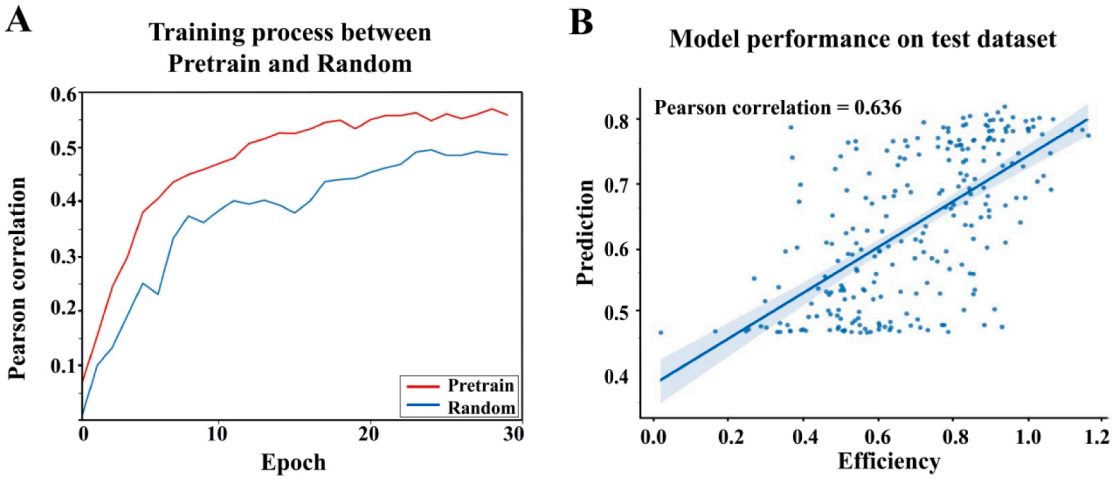

Fig. 3. Effect of the pretraining on the BERT-siRNA model. (A) Pearson correlation for the validation dataset during training, between pretraining and random initialization; (B) scatter plot of the prediction results for the well-trained pretraining model on the Huesken test dataset.

3.4. Model performance on experimental verification dataset

To further evaluate the robustness of our model, we retrieved the 11 potent siRNAs against SARS-CoV-2 (Chang et al., 2022). The siRNA with the prediction scores greater than 0.7 is classified as the high knockdown siRNA in our model. All the potent siRNAs are predicted as the high knockdown siRNAs in our model, the details are shown in Table 1.

In Fig. 4A, it can be observed that DISR is the best-performing model among the four on the independent test dataset. Therefore, we only select the DISR model for comparison with our model on our own laboratory data.

In our laboratory data, dataset1 consists of knockdown experiment data for *PDCD1*, dataset2 is for *CD38* and *IL6*, and dataset3 includes all data of three genes mentioned above.

The results are presented in Table 2. The correlation coefficients between the predicted outcomes and experimental results indicate that our BERT-siRNA model outperforms the existing DSIR model across different datasets.

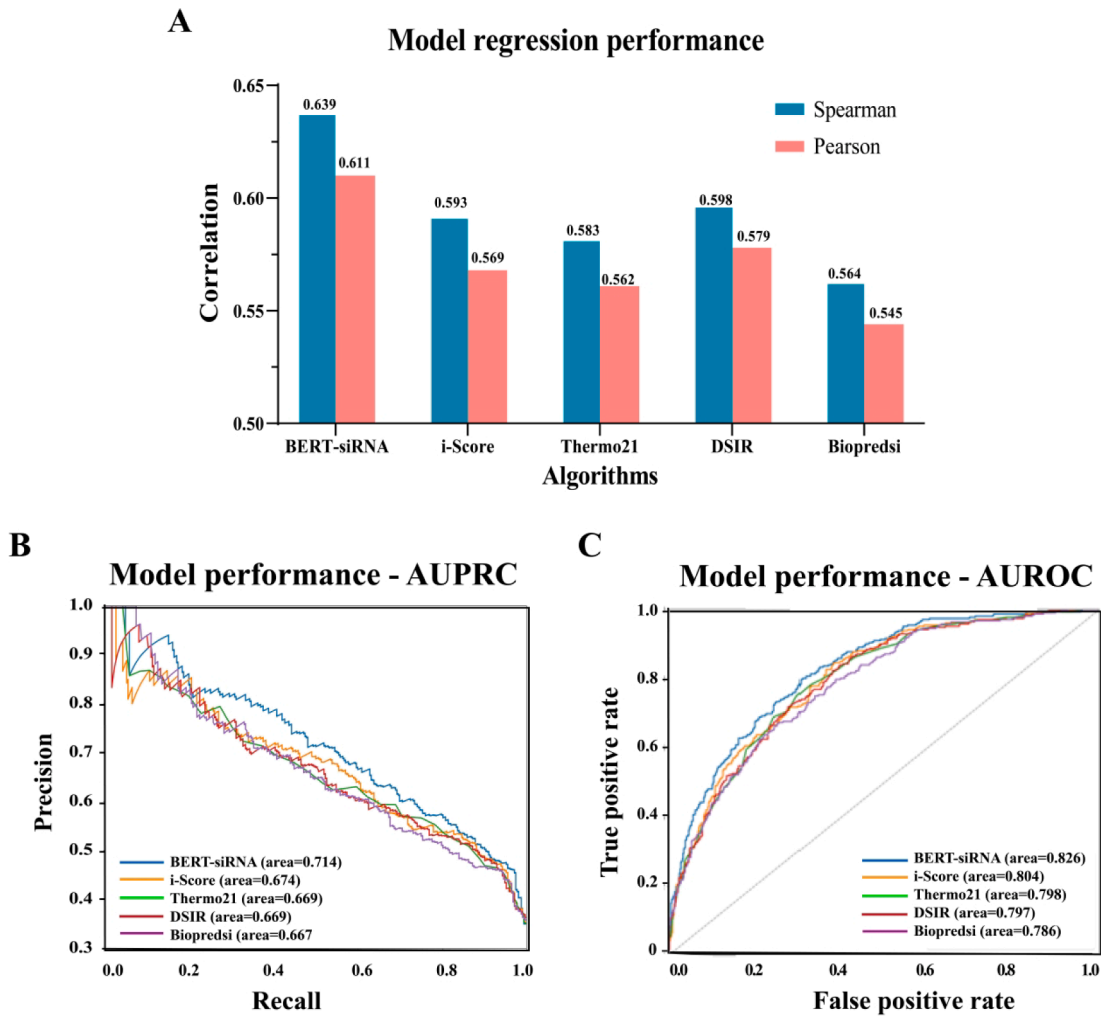

Fig. 4. Comparison with other siRNA prediction algorithms on the independent test dataset. (A) The Pearson and Spearman correlation to evaluate all the models' regression performance. (B) The AUPRC to evaluate all the models' classification performance. (C) The AUROC to evaluate all the models' classification performance.

Table 1
Model prediction for potent siRNA against SARS-Cov-2.

Name	Prediction score	High efficacy
C1	0.84	Y
C2	0.844	Y
C3	0.834	Y
C4	0.838	Y
C5	0.832	Y
C6	0.825	Y
C7	0.865	Y
C8	0.832	Y
C9	0.841	Y
C10	0.848	Y
C11	0.842	Y

Table 2
Comparison of prediction effect of DSIR model and BERT-siRNA model on knockdown laboratory datasets.

Experimental	DSIR	BERT-siRNA
Dataset 1	0.465	**0.57**
Dataset 2	0.494	**0.595**
Dataset 3	0.619	**0.669**

As shown in Fig. 5, we can see a clear trend that the siRNA with higher knockdown efficiency tends to have higher score in our model prediction, which the Pearson's correlation is 0.57 for *PDCD1*, 0.595 for *CD38* and *IL6*.

3.5. *Model explainability by visualization*

For a long time, the deep learning model has had the 'black-box' problem which deter us from better understanding and trusting the prediction. We think that the BERT model has the ability to discover the important pattern learned by the data which makes it a more explainable model. Therefore, we investigated the model's explainability through visualizing attention score and hidden layers to help us get a better understanding and put more trust on this model.

As shown in Fig. 6A, we first visualized the attention scores for all the independent test dataset through heatmap, and the calculation method is described in the Method part. Each row represents one siRNA and each column represents one nucleotide position. We can see that the attention scores are the highest in the 5′ end of the siRNA across all the dataset, which means that the nucleotides around the 5′ end influence the most in the model prediction. This finding is corresponding to the previous researches that the 5′ terminal nucleotides of siRNA/miRNA determinate the small RNA-binding affinity of AGO proteins and thus influence its biological activity (Frank et al., 2010; Mi et al., 2008; Wang et al., 2008; Yang et al., 2021). The attention score heatmap of one potent siRNA for

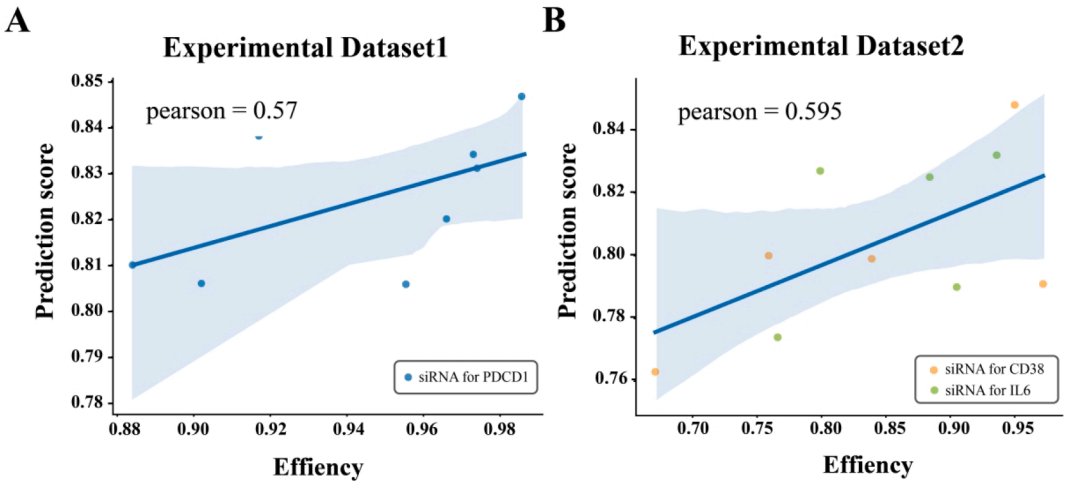

Fig. 5. Model prediction performance on laboratory siRNA knockdown data. (A) The result of knockdown experiments on *PDCD1* gene. (B) The result of knockdown experiments for *CD38* and *IL6* gene.

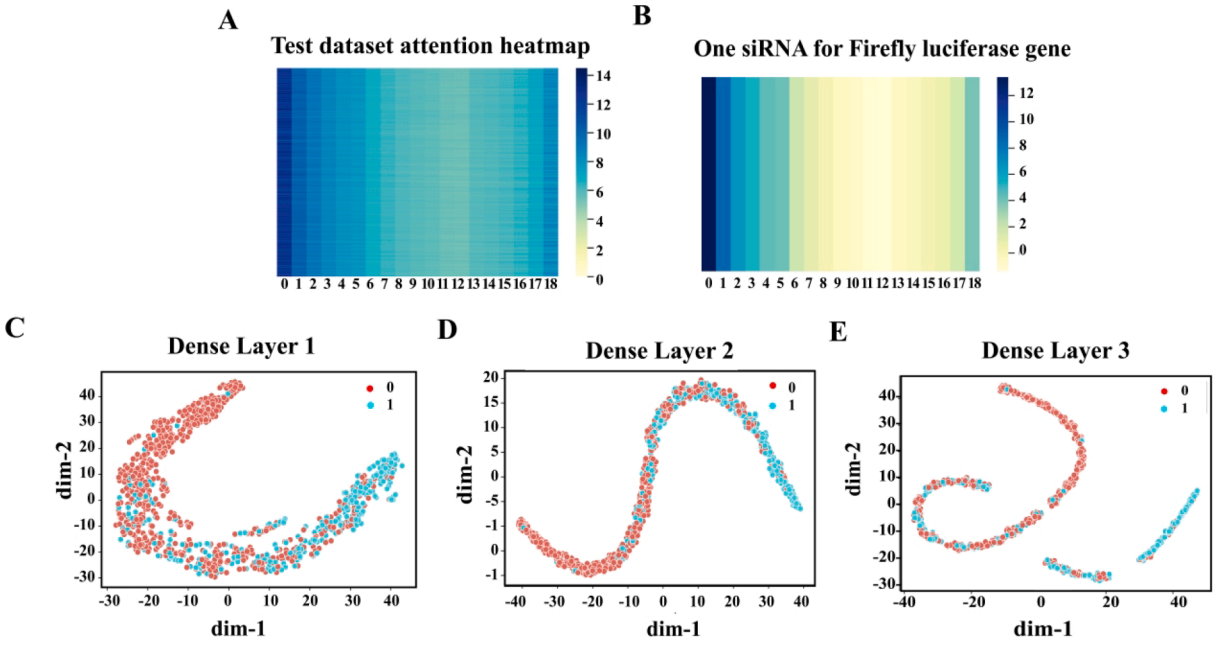

Fig. 6. BERT-siRNA explainability by visualization of the model parameters. (A) Attention heatmap for all the test dataset across 19 nucleotides. (B) Attention heatmap for one potent siRNA targeting Firefly luciferase gene. (C) TSNE plot for the hidden output of dense layer 1. (D) TSNE plot for the hidden output of dense layer 2. (E) TSNE plot for the hidden output of dense layer 3.

Fire luciferase gene is also shown in Fig. 5B, we can see a more obvious high value in the 5′ end of the siRNA.

We then extracted the hidden output of the final 3 dense layers to do the unsupervised clustering. These 3 layers have 128, 128 and 64 dimensions respectively. The clustering results are shown with the TSNE plot (Fig. 6C, D, E), the dataset is labeled by low and high knockdown groups (‘0′ for low and ‘1′ for high). We can see that the low and high knockdown siRNAs are separated gradually from intermediate to final dense layers. This finding explains the effective feature extraction of the MLP layers.

4. Discussion

With the outbreak of COVID-19, molecular therapy by siRNA is a prospective tool to control human viral infection by targeting conserved sequences of the virus. Compared with the long and complex shRNA vector (Zhao et al., 2022), the siRNA molecule is easy to be equipped and transduced into the cell, making it promising for drug development. siRNA target knockdown prediction is crucial for the mRNA inhibition, and several methods have been developed to enhance the prediction accuracy. Other siRNA methods have investigated the biological meaningful features to improve the prediction performance. For example, ThermoComposition21 used the thermodynamic and composition features, DSIR summarized a set of biological meaningful features and i-Score used nucleotide preferences at each position of siRNA as input to develop a linear regression model. However, prediction accuracy of current methods is far from our satisfactory. The main limitation that restricts us to get the accurate prediction is that we don’t have large enough dataset to train a perfect machine learning model. This limitation prevents the use of the state-of-the-art deep learning model which requires large dataset to optimize tens of thousands of parameters.

To overcome this problem, we used the BERT model which were

pretrained on the large relevant unlabeled dataset, and then finetuned it on the specific task. This strategy allows us to transfer the knowledge from larger similar dataset and enhance the deep learning performance for our task with relatively small dataset. To obtain a state-of-the-art deep learning model, we adopted DNABERT as the pretrained model and finetuned it on our siRNA prediction task. This is sensible since the siRNA sequences differ from DNA sequences only by one base, and most of the genomics text are similar. Further, the attention mechanism of the transformers-based model allows us to visualize the patterns learned by the attention head. The hidden output of the MLP intermediate layers could be visualized through clustering and explain the decision making of the network. These interpretable techniques make the BERT-siRNA an interpretable deep learning model, not just a 'black box'.

By comparing the model with or without pretraining by the DNA sequences, we found that the pretraining improved model performance. Thus, information that implies in the genome sequence is important for the siRNA target prediction. Our model could largely elevate the siRNA prediction performance compared with other siRNA methods on the large independent dataset, and perform better both in the regression or classification. Accurate prediction for siRNAs of SARS-CoV-2 shows high knockdown value. The siRNAs for *PDCD1*, *CD38* and *IL6* experiments in our lab further verified the robustness of our model. Nucleotides around siRNA 5′ end are the most important in our attention mechanisms, and this finding is corresponding to the previous siRNA experiment research. Insights on the rationale for deep learning model decisions are given by the high and low knockdown groups of the MLP hidden layers, which separate gradually as layers go deeper. Our model is solely based on the 19-nucleotide sequence, and no other extensive relevant biological features are used, making it applicable to various scenarios easily.

While, the limitations of this research should be acknowledged. Due to the lack of enough siRNA dataset, the powerful prediction ability of the deep learning model is still not fully released. Still, we provide an effective and robust tool for siRNA target design. The explainability of the model also provides us insight into the biological meaning and the inner mechanism of the BERT model.

5. Conclusions

This paper introduces a novel method for predicting siRNA target gene knockdown efficiency by leveraging a BERT model-based approach. The methodology employs transfer learning to incorporate sequence structure and dependency information from the genome into the model, enhancing prediction accuracy even with limited siRNA experimental knockdown data. The integration of visual analysis of model parameters enhances the biological explainability of the method, contributing to a comprehensive understanding of the predictive process.

Funding

This work was supported by China National Natural Science Foundation (62172121) and Natural Science Foundation of Heilongjiang Province of China (LH2022F012), and China National Natural Science Foundation (82073800).

Authors' contributions

WF and LY provided the idea and guidance. CZ and JX designed, implemented and wrote the manuscript. NX provided the biological insight and analysis. CZ provided the online service for the method. JX and WX helped do the biological validation analysis. All authors reviewed and approved the final version of the manuscript.

CRediT authorship contribution statement

Jiayu Xu: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Nan Xu:** Investigation, Data curation. **Weixin Xie:** Investigation. **Chengkui Zhao:** Writing – original draft, Visualization, Software, Conceptualization. **Lei Yu:** Resources. **Weixing Feng:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The codes and datasets are available online at https://github.com/ChengkuiZhao/siRNABERT.

Acknowledgements

The authors are grateful to acknowledge the fundings support provbided by the National Naturnal Science Foundation (62172121,82073800) and Natural Science Foundation of Heilongjiang Province of China (LH2022F012). In addition, we would like to thank Shanghai Unicar-Therapy Bio-medicine Technology Company for their support in biological experiments and EditSprings (https://www.editsprings.cn) for the expert linguistic services provided.

References

Chang, Y.C., Yang, C.F., Chen, Y.F., Yang, C.C., Chou, Y.L., Chou, H.W., Yang, P.C., 2022. A siRNA targets and inhibits a broad range of SARS-CoV-2 infections including Delta variant. EMBO Mol. Med. 14 (4) https://doi.org/10.15252/emmm.202115298.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint, arXiv: 1810.04805*.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C., 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391 (6669), 806–811. https://doi.org/10.1038/35888.

Frank, F., Sonenberg, N., Nagar, B., 2010. Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. Nature 465 (7299), 818–822. https://doi.org/10.1038/nature09039.

Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Hall, J., 2005. Design of a genome-wide siRNA library using an artificial neural network. Nat Biotechnol 23 (8), 995–1001. https://doi.org/10.1038/nbt1118.

Ichihara, M., Murakumo, Y., Masuda, A., Matsuura, T., Asai, N., Jijiwa, M., Ohno, K., 2007. Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. Nucleic Acids Res 35 (18), e123.

Izant, J.G., Weintraub, H., 1984. Inhibition of thymidine kinase gene expression by antisense RNA: a molecular approach to genetic analysis. Cell 36 (4), 1007–1015. https://doi.org/10.1016/0092-8674(84)90050-3.

Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V., 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 37 (15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083.

Katoh, T., Suzuki, T., 2007. Specific residues at every third position of siRNA shape its efficient RNAi activity. Nucleic Acids Res 35 (4), e27.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Qi, Y., 2008. Sorting of small RNAs into arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell 133 (1), 116–127. https://doi.org/10.1016/j.cell.2008.02.034.

Monty, M.A., Islam, M.A., Nan, X., Tan, J., Tuhin, I.J., Tang, X., Yu, L., 2021. Emerging role of RNA interference in immune cells engineering and its therapeutic synergism in immunotherapy. Br J Pharmacol 178 (8), 1741–1755. https://doi.org/10.1111/bph.15414.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A 116 (44), 22071–22080. https://doi.org/10.1073/pnas.1900654116.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., Khvorova, A., 2004. Rational siRNA design for RNA interference. Nat Biotechnol 22 (3), 326–330. https://doi.org/10.1038/nbt936.

Rossi, J.J., Rossi, D.J., 2021. siRNA drugs: here to stay. Mol Ther 29 (2), 431–432. https://doi.org/10.1016/j.ymthe.2021.01.015.

Saetrom, P., 2004. Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. Bioinformatics 20 (17), 3055–3063. https://doi.org/10.1093/bioinformatics/bth364.

Shabalina, S.A., Spiridonov, A.N., Ogurtsov, A.Y., 2006. Computational models with thermodynamic and composition features improve siRNA design. BMC Bioinf. 7, 65. https://doi.org/10.1186/1471-2105-7-65.

Vert, J.P., Foveau, N., Lajaunie, C., Vandenbrouck, Y., 2006. An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinf. 7, 520. https://doi.org/10.1186/1471-2105-7-520.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., Patel, D.J., 2008. Structure of the guide-strand-containing argonaute silencing complex. Nature 456 (7219), 209–213. https://doi.org/10.1038/nature07315.

Yang, P., Havecker, E., Bauer, M., Diehl, C., Hendrix, B., Hoffer, P., Deikman, J., 2021. Beyond identity: understanding the contribution of the 5' nucleotide of the antisense strand to RNAi activity. PLoS One 16 (9), e0256863.

Zhao, C., Xu, N., Tan, J., Cheng, Q., Xie, W., Xu, J., Feng, W., 2022. ILGBMSH: an interpretable classification model for the shRNA target prediction with ensemble learning algorithm. Brief Bioinform 23 (6). https://doi.org/10.1093/bib/bbac429.

Zhao, C., Cheng, Q., Xie, W., Xu, J., Xu, S., Wang, Y., Feng, W., 2022 May. Methods for predicting single-cell miRNA in breast cancer. Genomics 114 (3), 110353. https://doi.org/10.1016/j.ygeno.2022.110353.