NOVEL ENCODING OF SGRNA-DNA SEQUENCES FOR EFFECTIVE OFF-TARGET PREDICTIONS IN GENE EDITING WITH DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Off-target predictions are crucial in gene editing research to improve existing prediction methods. Recently, significant progress has been achieved in the field of prediction of off-target mutations, particularly with CRISPR-Cas9 data, thanks to the use of deep learning. CRISPR-Cas9 is a precise gene editing technique allowing manipulations of DNA fragments. The encoding of sgRNA-DNA sequences for deep neural networks is a complex process, which impacts significantly the prediction accuracy. In this context, we propose a novel encoding of sgRNA-DNA sequences that is capable to aggregate the involved sequence data without any loss of information. In our experiments, we compare our novel encoding with the state-of-the-art sgRNA-DNA encoding. We demonstrate the superior accuracy of our approach in our simulations involving Feedforward Neural Networks (FFNs) and Convolutional Neural Networks (CNNs). We highlight the universality of our results by building several FFNs and CNNs with various layer depths and performing predictions on two popular public gene editing data sets, the CRISPOR data set and the GUIDE-seq data set. In all our experiments, the new encoding led to more accurate off-target prediction results, providing an improvement of the area under the Receiver Operating Characteristic (ROC) curve up to 35%.

1 INTRODUCTION

Gene editing techniques allow genetic fragments to be added, removed or altered at specific locations of a genome Komor et al. (2017). A popular gene editing technique, called Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), has been recently developed Gupta & Musunuru (2014). CRISPR can be associated with the protein Cas9, and thus called CRISPR-Cas9. The protein Cas9 plays an important role in the immunological defense of certain bacteria. The bacteria capture fragments of the virus DeoxyriboNucleic Acid (DNA) and store them in CRISPR arrays for future virus attacks. In the case of a new virus attack, the bacteria use the protein Cas9 to produce RiboNucleic Acid (RNA) fragments from CRISPR arrays to cut the DNA of the virus and thus disable it. Researchers are able to use this technique to guide specific locations within complex genomes by a short RNA string Sander & Joung (2014). DNA sequences can therefore be easily edited or modulated in a variety of species and cell types, including human cell lines, bacteria, zebrafish and monkeys Sander & Joung (2014); Hsu et al. (2014). CRISPR-Cas9 technique has been extensively used for personalized therapy to edit and modulate harmful genes Kang et al. (2017); Liang et al. (2015). It was applied, for instance, in Ma et al. (2017) to correct a pathogenic mutation in a human embryos.

The targeting of specific DNA fragments in CRISPR-Cas9 is still, however, a challenging problem. The method relies on the Protospacer-Adjacent Motif (PAM) that dictates the DNA target search mechanism. The PAM is located at the end of the DNA target site. Shah et al.Shah et al. (2013) suggested that the PAM recognition is responsible for the relationship between target binding and cleavage conformations Hsu et al. (2014). It has been nonetheless observed that Cas9-sgRNA-DNA can target other DNA fragments than the original DNA fragment aimed, leading to off-target mutations Chen et al. (2017). Kim et al.Kim et al. (2015); Zhang et al. (2015) showed that CRISPR-Cas9 can tolerate mismatches in sgRNA-DNA at different locations, hence contributing to off-target mutations. The latter may provoke genomic instability and modify the gene behavior. Clinical ap-

plications still face this important challenge of the CRISPR-Cas9 gene editing method. Researchers are consequently working on off-target prediction methods to help isolating the exact location of the DNA cleavage.

Recent off-target prediction methods employ different scores highlighting the locations of the mismatches Haeussler et al. (2016); Xu et al. (2017). Popular off-target prediction scores include Cutting Frequently Determination (CFD), CROP-IT, CCTop and MIT. CFD score is evaluated based on the percentage activity rates of different mutations by simulating sgRNAs modifications with reference to validated sgRNAs Doench et al. (2016); Lin & Wong (2018). The CROP-IT score Singh et al. (2015) is determined by dividing each DNA fragment into sub-fragments with different weights, allowing one to grade off-target sgRNA sequences. CCTop Stemmer et al. (2015) and MIT Hsu et al. (2014) scores count the mismatched sgRNA-DNA sites to detect potential off-targets. As mentioned in Lin and Wong Lin & Wong (2018), these scoring methods: (i) do not take advantage of the growing CRISPR-Cas9 data sets, and (ii) do not establish the relationship between mismatched and matched sites. Lin and Wong Lin & Wong (2018) proposed to use deep learning for off-target predictions in CRISPR-Cas9 to solve the shortcomings of traditional scoring methods. Each genomic sequence is considered as an input for a deep neural network using a specific encoding of 4 nucleotides, Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), present in a DNA sequence. However, as we will demonstrate it in our experiments, the encoding proposed by Lin and Wong Lin & Wong (2018) aggregates the sgRNA and DNA nucleotide data losing some important information and thus impacting the performance of the off-target predictions. In this paper, we consider the genomic sequences as a $(2 \times 4) \times L$ matrix, where 2 reflects the duality between the sgRNA and DNA, 4 is the number of different nucleotides, and L is the length of the sequence-pair. Our main contributions are summarized below:

- We propose a novel sequence encoding technique that maps each sgRNA-DNA sequence pair into a $8 \times L$ matrix. The matrix of the sequence pairs is used as input of the deep learning models for the off-target predictions in CRISPR-Cas9 gene editing.
- We propose and test different architectures of deep neural networks, including feedforward neural networks and convolutional neural networks designed and optimized for off-target predictions on two well-known gene editing data sets, CRISPOR and GUIDE-seq.
- We demonstrate that the use of our encoding allows one to outperform the state-of-the-art off-target prediction methods, for all types of deep neural networks considered and for both CRISPOR and GUIDE-seq data sets.

The paper is structured as follows. We discuss the related work in Section 2. We then describe in Section 3 how we encode each sgRNA-DNA sequence pair into a $8 \times L$ matrix. We underline that the $8 \times L$ encoding is at the core of our approach, preventing any loss of information that may occur with other types of encoding. In Section 4, we show that the use of our encoding allows us to achieve a superior off-target prediction accuracy for different types of deep neural networks and different experimental data sets. Finally, we conclude and address some promising directions for future work.

2 Related Work

Recently, the deep learning approachGoodfellow et al. (2016) has become extremely popular thanks to multiple effective applications in different fields, including bioinformatics and health science. In genome editing, popular neural network models are Convolutional Neural Networks (CNNs) Fukushima et al. (1983); Lang et al. (1990) and Feedforward Neural Networks (FNNs) Hinton (1990). CNNs are extensively used in image recognition and vision computing. Nonetheless, CNNs can be also adapted to genome editing by considering each sgRNA-DNA sequence as a black and white image since sgRNA-DNA sequences can be stored in a matrix Lin & Wong (2018). Other popular machine learning methods such as Random Forest (RF) Breiman (2001) or Support Vector Machine (SVM) Platt et al. (1999); Chang & Lin (2011) have been as well used in genome editing Jarquín et al. (2014); Sheng et al. (2007). These techniques have however become less popular nowadays in favor of deep learning, due to the complexity and the constant increasing volume of genomic data sets Lin & Wong (2018).

Cas9 still requires a better understanding of its specificity for on- and off- target DNA binding and cleavage profiles with further studies Hsu et al. (2014). Wu et al.Wu et al. (2014) proposed Cas9-based chromatin immunoprecipitation sequencing (ChIP-seq) analysis for a better understanding of the binding degeneracy. Crosetto et al.Crosetto et al. (2013) addressed the problem of construction of a map of Cas9-induced off-targets. These studies are the premises of the generation of the data sets that will help create predictive models aiming at minimizing the off-target effect in genome editing applications. Tsai et al. Tsai et al. (2015) proposed GUIDE-seq to profile off-target cleavage by CRISPR-Cas9 nucleases. In their framework, the authors highlighted off-targets that can be characterized as validated off-targets or putative off-targets, delivering a unique opportunity to build prediction models using the traditional machine learning and deep learning approaches. It is worth noting however that the sample size of the GUIDE-seq data set is limited, minimizing the possibility of building accurate off-target predictive methods. Haeussler et al. Haeussler et al. (2016) addressed these needs by collecting data from different experimental studies Tsai et al. (2015); Hsu et al. (2013); Ran et al. (2015); Kim et al. (2016); Frock et al. (2015); Cho et al. (2014). These researchers were able to gather a large data set containing more than 26,000 of validated and putative off-targets. This volume of genetic data allows one to build off-target prediction models using traditional machine learning and deep learning models as the sample size became statistically relevant. Peng et al. Peng et al. (2018) proposed an ensemble support vector machine method for the prediction of off-target sites of sgRNA sequence data. By using on-target site sequences of sgRNA and genome-wide candidates for off-target sites, these authors predicted the labels of candidate for off-target site as real off-target if the predicted score was superior to a fixed threshold. Lin and Wong Lin & Wong (2018) proposed to use CNNs and FNNs for predicting off-target mutations. The last two papers establish the foundation of the use of machine learning and deep learning models in the field of CRISPR-Cas9, contributing to the need of a better understanding of the on- and off-target DNA binding. It remains a strong requirement for clinical applications in genome editing.

3 Proposed Method

3.1 Background on Nucleobases, guide RNA and PAM

In the genome editing with the CRISPR method, two components are required: a guide RNA and a CRISPR-associated endonuclease (a Cas protein sequence) McDade (2017). A sequence of around 20 nucleotide spacers defines the genomic target, or target DNA, to be altered. We recall that four nucleobases can be present in a DNA sequence: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The four nucleobases of a RNA sequence are: Adenine (A), Cytosine (C), Guanine (G) and Uracil (U), replacing Thymine (T). The sequence of nucleobases can be either a DNA or a RNA sequence. The 20 nucleotide spacers sequence is unique compared to the rest of the genome, and the target DNA is adjacent to a Protospacer Adjacent Motif (PAM). The latter is a short three-to-five nucleobase sequence serving as the binding signal of the Cas protein. However, the target sequence has very likely a homology sequence elsewhere in the genome, leading to undesired off-target sequences. This underlines the need for the development of off-target prediction methods.

3.2 Encoding sgRNA-DNA as $8 \times L$ matrix

The core of our contribution is our novel encoding of the sgRNA-DNA as a matrix of size $8 \times L$. As recalled in the previous sub-section, the DNA and the sgRNA sequences can include four different nucleobases. Each of these nucleobases is one-hot encoded. For the DNA, the four nucleobases, Adenine, Guanine, Cytosine and Thymine, can be represented as the following one-dimensional vectors: $[1, 0, 0, 0]$, $[0, 1, 0, 0]$, $[0, 0, 1, 0]$ and $[0, 0, 0, 1]$. The four nucleobases of the RNA, Adenine, Guanine, Cytosine and Uracil, can be similarly encoded as the following one-hot encoded vectors: $[1, 0, 0, 0]$, $[0, 1, 0, 0]$, $[0, 0, 1, 0]$ and $[0, 0, 0, 1]$. In the data sets used in our experiments, the sgRNA sequences have already been converted into the DNA sequences, meaning that Uracil was replaced Thymine. This had no impact on our results since the one-hot vector encoding was preserved. This way, a complete nucleobase sequence of length L can be one-hot encoded. This operation results in a matrix of size $4 \times L$ representing the DNA nucleobase sequence and a matrix of size $4 \times L$ representing the sgRNA nucleobase sequence. The two sub-matrices are then appended together to form a matrix of size $8 \times L$. The one-hot encoding is carried out similarly for all the sgRNA-DNA nucleobase sequences available in a given data sets.

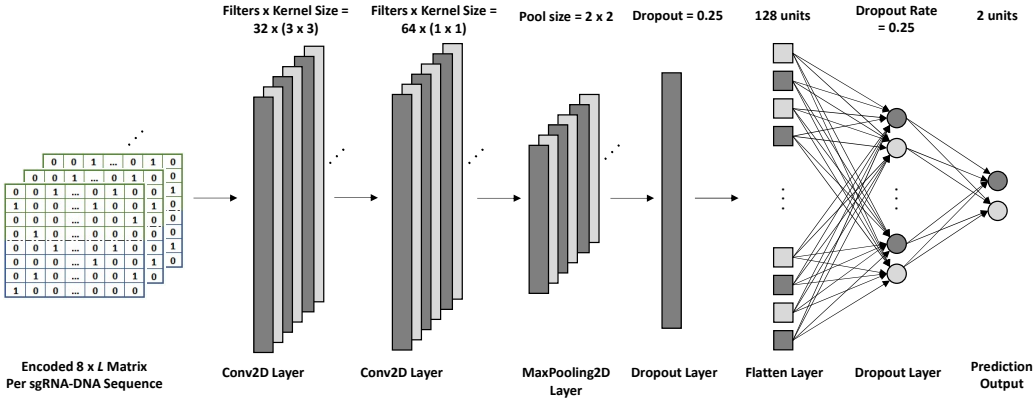


Figure 1: Representation of a standard architecture of a CNN used for off-target predictions. The encoded matrix containing the sgRNA-DNA information of each nucleobase sequences is used as input. Different convolutional layers capture the sgRNA-DNA information of the encoded matrix, a maxpooling is used to downsample the output of the convolutional layers. Fully connected layers are finally used to perform off-target predictions. Minor variations exist between different CNN architectures used in our experiments depending on the number of layers considered.

Lin and Wong Lin & Wong (2018) proposed to use a complementary base to represent the original base in the sgRNA sequences. Such a complementary base allows one to employ the A, G, C and T nucleobases to represent both the sgRNA and the target DNA sequence in CRISPR (Cas9). Encoding each base in the sgRNA and the target DNA sequences as one of the four one-hot vectors, Lin and Wong represented every sgRNA-DNA sequence pair by a $4 \times L$ matrix. We would like to highlight that contrary to the encoding of Lin and Wong, our encoding does not require consideration of the complementary base in the sgRNA sequence. The encoding we propose is much more flexible and prevent any information loss as the sgRNA and the DNA nucleobase sequences remain independent during the encoding process and in the resulting matrix. The $8 \times L$ matrix is finally used as input for the neural network predictions.

3.3 Neural Networks for Off-Target Predictions

We define two types of deep neural networks, CNNs and FFNs for off-target predictions.

In our CNN architecture, the matrix of size $8 \times L$, containing the sgRNA-DNA information is convoluted by a layer with a kernel size of 3×3 using a Rectified Linear Unit (ReLU) activation Glorot et al. (2011). The information from the first convolutional layer is then passed into the second convolutional layer with a kernel of size 1×1 with ReLU activation. The first and the second layers are designed to extract the sgRNA-DNA information. The third layer is a MaxPooling layer with a pool size equal to $(2, 2)$. Such an architecture allows us to evaluate the largest value of each sub-region. We then flatten the output of the MaxPooling layer to connect it to a dense layer with 128 neurons with a ReLU activation. A dropout layer is used as a regularization method to prevent overfitting Srivastava et al. (2014). Finally, the last fully connected layer carries out the predictions of the off-targets with a softmax activation. Minor variations exist between the different CNN architectures used in our experiments, depending on the number of layers considered.

FNN is the second type of deep neural networks used in our experiments. The FNN architecture is slightly less complex than the CNN architecture, but it performed equally well in our simulations. The $8 \times L$ matrix with the sgRNA-DNA information is used as input. A first dense layer of 100 neurons collects the information from the sgRNA-DNA matrix. A batch normalization layer Ioffe & Szegedy (2015) is fully connected to the first layer. A second dense layer with 75 neurons is used, followed by a batch normalization. A third dense layer with 50 neurons is applied with a dropout layer at the output. Two more fully connected layers with 25 and 10 neurons are used before reaching the last layer responsible for the off-target predictions. All the dense layers have a uniform kernel initializer and a ReLU activation unit, except for the last layer which has a softmax activation.

Table 1: Taxonomy of the models used in our experiments and their respective architecture description.

Type	Name	Architecture
FNN	FNN 3 Layers (FNN 3)	3 fully connected dense layers
FNN	FNN 5 Layers (FNN 5)	5 fully connected dense layers
FNN	FNN 10 Layers (FNN 10)	10 fully connected dense layers
CNN	CNN 3 Layers (CNN 3)	1 convolutional layer, 2 fully connected dense layers
CNN	CNN 5 Layers (CNN 5)	2 convolutional layer, 3 fully connected dense layers
CNN	CNN Lin et al. Wong	CNN model Lin & Wong (2018) from Lin and Wong GitHub page

Table 1 reports some details regarding the architecture of each model used in our experiments. Further information concerning the implementation of our models is available in our GitHub repository Anonymous (2020).

4 EXPERIMENTS

4.1 DATA AVAILABILITY

Two well-known CRISPR data sets were considered in our experiments: the CRISPOR data set Haeussler et al. (2016) and the GUIDE-seq data set Tsai et al. (2015). Haeussler et al. Haeussler et al. (2016) collected the CRISPOR data by gathering information from eight CRIPSR-Cas9 off-target studies. They built a dedicated website `http://crispor.tefor.net`, now established as a reference, highlighting their experiments and results. The CRISPOR online database contains 526 genomes, as of May 2020, including the human genome and the genomes of different plants, bacteria and viruses. The data base keeps growing with more and more genomes being available. In the data set used in the experiments, 26,052 putative off-targets have been identified using 177 validated off-targets, label as 1 representing the positive class. We refer the reader to the CRISPOR online documentation available at `http://crispor.tefor.net` for further details regarding the on-target and off-target predictions with CRISPOR data. The GUIDE-seq data set was published in supplementary material of Tsai et al. (2015). The GUIDE-seq method applies the CRISPR RNA-guided nucleases (RGNs) to two human cell lines, U2OS and HEK293 site 2. Tsai et al.Tsai et al. (2015) considered different sites in their experiments, such as VEGFA sites 1, 2 and 3 or HEK293 sites 2, 3 and 4. Among the 442 sgRNA-DNA sequence-pairs available in GUIDE-seq, 30 are validated off-targets, labeled as 1 and representing the positive class. For both data sets, the non-validated off-targets are labeled as 0, representing the negative class.

4.2 NEURAL NETWORKS MODELS COMPARED IN OUR EXPERIMENTS

In our experiments, we relied on different neural network architectures to assess the impact of the model on our encoding. We designed neural network models with various numbers of layers. We refer to Table 1 for a short architecture description of the neural networks used in our study. We limited ourselves to the use of 10 layers maximum for the FNNs. Precisely, we implemented three FNN models with 3, 5 and 10 layers. We also proposed and tested two CNN models with 3 and 5 layers. We moreover found and tested the CNN model of Lin and Wong Lin & Wong (2018), called Lin et al. in our experiments, on the authors GitHub page. It was used as benchmark to validate the results of our experiments.

4.3 EXPERIMENTAL PROCEDURE AND TRAIN-TEST SPLIT

We divided our experiments into two main parts. We first train and test the models with the CRISPOR data set, and then we apply transfer learning on the GUIDE-seq data set.

The original CRISPOR data set was split into two sub-sets, a training data set and a test data set. We used a standard train-test split from the *sklearn* implementation with shuffling, a ratio of 0.3 and equal stratification of the classes. As we were not performing any features selection, the stratified k-fold validation was not required. We trained our prediction models on the CRISPOR training data set and cross-validated the predictions on the independent CRISPOR test data set (the data sets are available at our GitHub repository). Our train-test split satisfies two requirements to ensure that the training and test sets are independent and do not have identical examples: (i) the nucleobase

Table 2: Performance metrics assessed for each predictive model used with 4×23 and 8×23 encodings on the CRISPOR data set. These performance metrics highlight the superior accuracy of the predictive models using the 8×23 encoding.

Metric	Encoding	FNN 3	FNN 5	FNN 10	CNN 3	CNN 5	CNN Linn et al.	RF
Mean Accuracy	4×23	0.995	0.995	0.994	0.995	0.995	0.894	0.995
Mean Accuracy	8×23	0.999	0.999	0.999	0.995	0.997	0.994	0.998
Brier Score	4×23	0.004	0.004	0.004	0.005	0.005	0.216	0.004
Brier Score	8×23	0.001	0.001	0.001	0.004	0.003	0.160	0.001
Precision Score	4×23	1.000	0.652	0.000	0.833	0.625	0.010	0.750
Precision Score	8×23	1.000	1.000	1.000	1.000	1.000	0.994	1.000
Recall Score	4×23	0.140	0.349	0.000	0.116	0.116	0.186	0.209
Recall Score	8×23	0.791	0.860	0.884	0.116	0.442	1.000	0.721
F1 Score	4×23	0.245	0.455	0.000	0.204	0.196	0.020	0.327
F1 Score	8×23	0.883	0.923	0.938	0.208	0.613	0.020	0.838
AUC ROC	4×23	0.948	0.925	0.960	0.929	0.913	0.730	0.907
AUC ROC	8×23	0.984	0.995	0.975	0.952	0.968	0.780	0.984
AUC PR 1	4×23	0.390	0.319	0.439	0.287	0.253	0.016	0.393
AUC PR 1	8×23	0.891	0.949	0.936	0.684	0.785	0.027	0.909

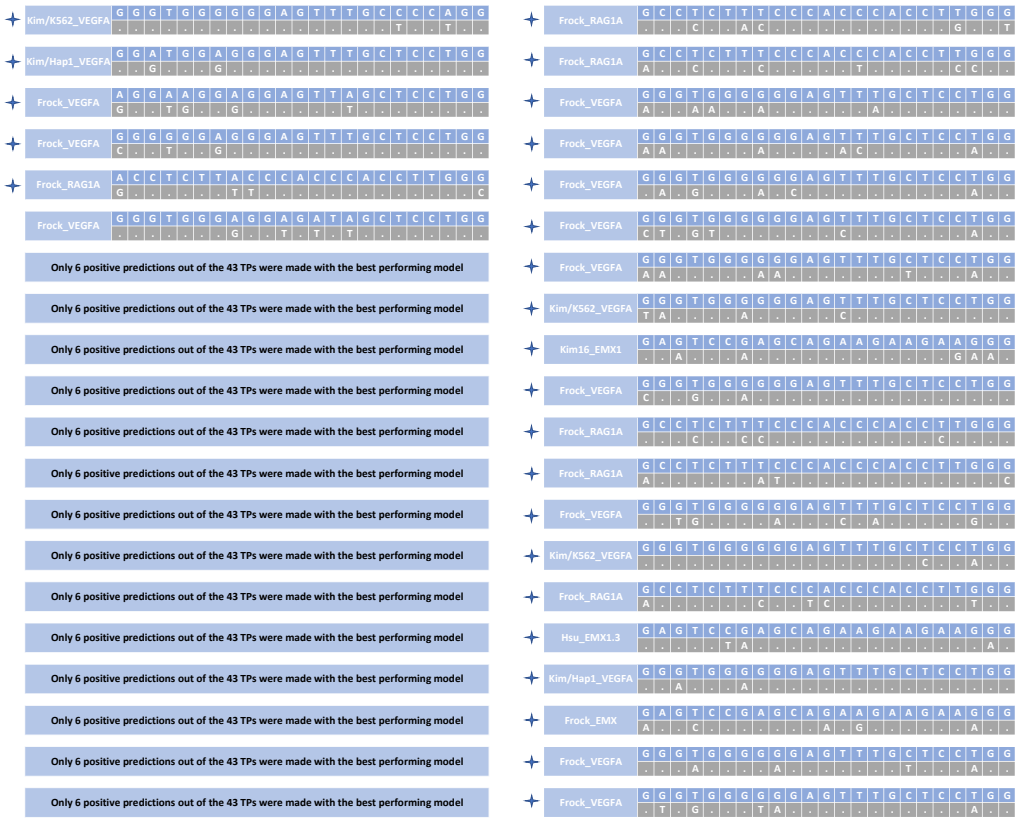
sequence pairs in the whole data set are all unique, meaning that there are no homologous nucleobase sequence pairs in the training and test sets, and (ii) the average pairwise matrix similarity, calculated using the 2-norm (**?**) between the matrices encoding sgRNA-DNA sequence pairs (see Figure 1), within the training set, within the test set, and between the training and test sets are very similar, 89.56%, 89.42% and 89.44%, respectively (high values of the matrix norm are due to the high sparsity of the matrices).

The GUIDE-seq data set has a limited number of samples, containing only 430 unique nucleobase sequence pairs. All of them are different from the CRISPOR training and test data sets. Thus, the data sets used for training (CRISPOR data set) and cross-validation (GUIDE-seq data set) were completely independent in this experiment. As the number of samples in GUIDE-seq is too small to build training and test data sets for deep learning predictions, we applied the transfer learning approach (Caruana, 1997) to perform off-target predictions on this data. In transfer learning, a model trained on one data set for a specific application is then used to perform predictions on a different data set for a similar application. We consequently used the trained models of the CRISPOR experiment for off-target predictions on the GUIDE-seq data set.

4.4 Results and Discussion

We begin by discussing the results of the first experiment - the off-target prediction on the CRISPOR data set to which a train-test split has been applied. We present quantitative results in Table 2. For all models, the quantitative metrics show gains in accuracy of off-target predictions, which is often very important, provided by the new encoding. The Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) for all tested neural networks and the Random Forest (RF) model, the encoding 8×23 leads to greater values than the 4×23 encoding. Similarly, the comparison of the AUC PR1 statistic for the two competing encodings shows that the 8×23 encoding leads to much higher values of AUC PR1. We present the top 20 predictions of the models on the CRISPOR data set in Figure 2. We took the predicted positive samples and ranked them in the descending order of probability. In Figure 2a, six positive predictions were made by the FNN with 5 layers for the encoding 4×23. The FNN with 5 layers is the best performing model according to F1-score. Among the six positive predictions, five were true positives and only one was false positive. However, the same FNN with 5 layers could only predict six positive samples out of 43 true positives present in the test set when the 4×23 encoding was used. In contrast, as shown in Figure 2b, the top 20 predictions obtained with the 8×23 encoding by the FNN with 10 layers (the best F1-score performing model for this encoding) were all true positives. Thus, Figure 2 strongly emphasizes the superior accuracy of the positive class predictions obtained using the most accurate deep learning models and the proposed 8×23 data encoding. By aggregating the results presented in Table 2 and in Figure 2a, we can conclude that our novel 8×23 encoding leads to the superior accuracy of the off-target predictions, independently of the machine learning model being used.

We describe hereinafter the results of our second experiment carried out with the GUIDE-seq data set of a much smaller size. We therefore applied transfer learning using machine learning models trained on CRISPOR data to perform the predictions of the GUIDE-seq off-targets. Table 3 presents



(a) Top 20 predictions obtained using the 4×23 encoding with the FNN with 5 layers.

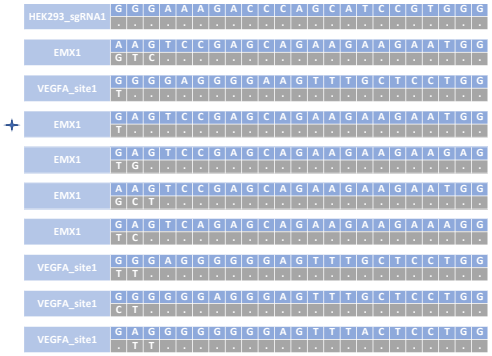
(b) Top 20 predictions obtained using the 8×23 encoding with the FNN with 10 layers.

Figure 2: Top 20 off-target predictions obtained for the CRISPOR data set using the most accurate deep learning model (according to F1-score) for: (a) the 4×23 and (b) the 8×23 sequence encodings. The sgRNA-DNA sequences marked by a star are true positive off-targets, and false positive off-targets otherwise. The top 20 predictions obtained with the 8×23 encoding were all true positives. The best performing model with the 4×23 encoding could only predict 6 positive samples, among which 5 were true positive samples and 1 was a false positive sample.

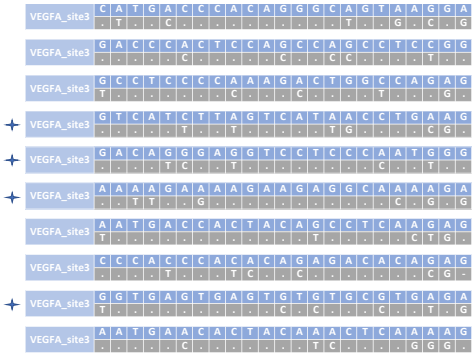
a quantitative overview of the performance of both encodings in the framework of transfer learning. The 8×23 encoding shows an important gain in accuracy in the predictions. The difference provided by the two competing encodings was greater in this experiment with an improvement of the ROC AUC value up to 35% (for the FNN with 5 layers). The AUC PR1 statistic underlines the superior accuracy of the 8×23 encoding for all the models except for the Lin and Wong CNN model Lin & Wong (2018). Surprisingly, this model did not perform as well as the others with the transfer learning. It was the worst performing model in our experiments. Furthermore, in Figure 3 we present the top 10 predictions obtained using the best F1-score performing models on the GUIDE-seq data set. The best F1-score performing model for the 4×23 encoding was the FNN with 5 layers, whereas for the 8×23 encoding it was the FNN with 3 layers. Similarly to what we did in the first experiment, we ranked by descending order of probability the predicted positive samples. In Figure 3a, only one true positive off-target was predicted using the 4×23 encoding. Four true positives were predicted using the 8×23 encoding in Figure 3b. Thus, Figure 3 highlights the superior accuracy of the off-target predictions obtained with the 8×23 encoding with respect to the best F1-score performing model. Albeit the transfer learning approach does have an impact on the predictions accuracy, the results presented in Table 3 and in Figure 3a still underline the superior accuracy of the off-target predictions obtained with the 8×23 data encoding.

Table 3: Performance metrics assessed for each predictive model used with 4×23 and 8×23 encodings on the GUIDE-seq data set. These performance metrics highlight the superior accuracy of the predictive models using the 8×23 encoding.

Metric	Encoding	FNN 3	FNN 5	FNN 10	CNN 3	CNN 5	CNN Linn et al.	RF
Mean Accuracy	4×23	0.936	0.919	0.945	0.948	0.948	0.817	0.910
Mean Accuracy	8×23	0.903	0.889	0.932	0.932	0.932	0.932	0.916
Brier Score	4×23	0.053	0.056	0.052	0.047	0.048	0.225	0.069
Brier Score	8×23	0.059	0.065	0.061	0.064	0.064	0.175	0.062
Precision Score	4×23	0.250	0.238	0.000	0.667	0.667	0.000	0.000
Precision Score	8×23	0.303	0.148	0.000	0.000	0.000	0.000	0.000
Recall Score	4×23	0.087	0.217	0.000	0.087	0.087	0.087	0.000
Recall Score	8×23	0.333	0.133	0.000	0.000	0.000	0.000	0.000
F1 Score	4×23	0.129	0.227	0.000	0.154	0.154	0.049	0.000
F1 Score	8×23	0.317	0.140	0.000	0.000	0.000	0.000	0.000
AUC ROC	4×23	0.615	0.511	0.579	0.756	0.629	0.508	0.566
AUC ROC	8×23	0.899	0.863	0.859	0.881	0.851	0.545	0.833
AUC PR 1	4×23	0.133	0.132	0.191	0.279	0.246	0.166	0.091
AUC PR 1	8×23	0.282	0.214	0.296	0.376	0.325	0.068	0.196



(a) Top ten predictions obtained using the 4×23 encoding with the FNN with 5 layers.



(b) Top ten predictions obtained using the 8×23 encoding with the FNN with 3 layers.

Figure 3: Top 10 off-target predictions obtained for the GUIDE-seq data set using the most accurate deep learning model (according to F1-score) for: (a) the 4×23 and (b) the 8×23 sequence encodings. The sgRNA-DNA sequences marked by a star are true positive off-targets, and false positive off-targets otherwise. Four of the top ten predictions obtained with the 8×23 encoding are true positives. The best performing model with the 4×23 encoding could only predict one true positive.

5 Conclusion

In this paper, we introduced a novel sgRNA-DNA sequence encoding technique for effective off-target predictions with deep neural networks. We proposed to encode the nucleobase sequence pairs as a matrix of size 8×23, allowing the storage of both sgRNA and DNA information independently. Our approach prevents any loss of genetic information during the encoding process. We performed our experiments on two popular sgRNA-DNA data sets, i.e. the CRISPOR data set and the GUIDE-seq data set. Different neural network models were implemented and tested in this work to show the universality of the proposed approach and the consistency of our results. In our experiments, we demonstrated that the proposed encoding was capable to consistently outperform the current state-of-the-art sgRNA-DNA encoding, consisting in a matrix representation of size 4×23. The gain in performance provided by the 8×23 encoding led to a significant improvement of all evaluation metrics, independently of the machine learning model considered. Overall, our approach leads to much better predictions of the off-target mutations in genome editing, aiming ultimately at attaining the precision requirements of clinical applications intended for genome editing. Our future work will address the off-target predictions of CRISPR-Cas12, for which the target binding is less sensitive to the nucleobases sequence. We will focus as well on the impact of the off-targets class imbalance. We will finally investigate possible improvements of sgRNA-DNA encoding for deep learning using a three-dimensional nucleobase sequence-pairs encoding.

REFERENCES

Anonymous. Reference removed to preserve the anonymity of the reviewing process. Reference Removed to Preserve the Anonymity of the Reviewing Process, 2020.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Janice S Chen, Yavuz S Dagdas, Benjamin P Kleinstiver, Moira M Welch, Alexander A Sousa, Lucas B Harrington, Samuel H Sternberg, J Keith Joung, Ahmet Yildiz, and Jennifer A Doudna. Enhanced proofreading governs crispr–cas9 targeting accuracy. *Nature*, 550(7676):407–410, 2017.

Seung Woo Cho, Sojung Kim, Yongsub Kim, Jiyeon Kweon, Heon Seok Kim, Sangsu Bae, and Jin-Soo Kim. Analysis of off-target effects of crispr/cas-derived rna-guided endonucleases and nickases. *Genome research*, 24(1):132–141, 2014.

Nicola Crosetto, Abhishek Mitra, Maria Joao Silva, Magda Bienko, Norbert Dojer, Qi Wang, Elif Karaca, Roberto Chiarle, Magdalena Skrzypczak, Krzysztof Ginalski, et al. Nucleotide-resolution dna double-strand break mapping by next-generation sequencing. *Nature methods*, 10(4):361, 2013.

John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2): 184, 2016.

Richard L Frock, Jiazhi Hu, Robin M Meyers, Yu-Jui Ho, Erina Kii, and Frederick W Alt. Genome-wide detection of dna double-stranded breaks induced by engineered nucleases. *Nature biotechnology*, 33(2):179, 2015.

Kunihiko Fukushima, Sei Miyake, and Takayuki Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5):826–834, 1983.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, 2011.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Rajat M Gupta and Kiran Musunuru. Expanding the genetic editing tool kit: Zfns, talens, and crispr-cas9. *The Journal of clinical investigation*, 124(10):4154–4161, 2014.

Maximilian Haeussler, Kai Schönig, Hélène Eckert, Alexis Eschstruth, Joffrey Mianné, Jean-Baptiste Renaud, Sylvie Schneider-Maunoury, Alena Shkumatava, Lydia Teboul, Jim Kent, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide rna selection tool crispor. *Genome biology*, 17(1):148, 2016.

Geoffrey E Hinton. Connectionist learning procedures. In *Machine learning*, pp. 555–610. Elsevier, 1990.

Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, 31(9):827, 2013.

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

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Diego Jarquín, Kyle Kocak, Luis Posadas, Katie Hyma, Joseph Jedlicka, George Graef, and Aaron Lorenz. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC genomics*, 15(1):740, 2014.

Xiang Jin Kang, Chiong Isabella Noelle Caparas, Boon Seng Soh, and Yong Fan. Addressing challenges in the clinical applications associated with crispr/cas9 technology and ethical questions to prevent its misuse. *Protein & cell*, 8(11):791–795, 2017.

Daesik Kim, Sangsu Bae, Jeongbin Park, Eunji Kim, Seokjoong Kim, Hye Ryeong Yu, Jinha Hwang, Jong-Il Kim, and Jin-Soo Kim. Digenome-seq: genome-wide profiling of crispr-cas9 off-target effects in human cells. *Nature methods*, 12(3):237, 2015.

Daesik Kim, Sojung Kim, Sunghyun Kim, Jeongbin Park, and Jin-Soo Kim. Genome-wide target specificities of crispr-cas9 nucleases revealed by multiplex digenome-seq. *Genome research*, 26 (3):406–415, 2016.

Alexis C Komor, Ahmed H Badran, and David R Liu. Crispr-based technologies for the manipulation of eukaryotic genomes. *Cell*, 168(1-2):20–36, 2017.

Kevin J Lang, Alex H Waibel, and Geoffrey E Hinton. A time-delay neural network architecture for isolated word recognition. *Neural networks*, 3(1):23–43, 1990.

Puping Liang, Yanwen Xu, Xiya Zhang, Chenhui Ding, Rui Huang, Zhen Zhang, Jie Lv, Xiaowei Xie, Yuxi Chen, Yujing Li, et al. Crispr/cas9-mediated gene editing in human tripronuclear zygotes. *Protein & cell*, 6(5):363–372, 2015.

Jiecong Lin and Ka-Chun Wong. Off-target predictions in crispr-cas9 gene editing using deep learning. *Bioinformatics*, 34(17):i656–i663, 2018.

Hong Ma, Nuria Marti-Gutierrez, Sang-Wook Park, Jun Wu, Yeonmi Lee, Keiichiro Suzuki, Amy Koski, Dongmei Ji, Tomonari Hayama, Riffat Ahmed, et al. Correction of a pathogenic gene mutation in human embryos. *Nature*, 548(7668):413–419, 2017.

Joel McDade. The pam requirement and expanding crispr beyond spcas9. *Addgene*, 2, 2017.

Hui Peng, Yi Zheng, Zhixun Zhao, Tao Liu, and Jinyan Li. Recognition of crispr/cas9 off-target sites through ensemble learning of uneven mismatch distributions. *Bioinformatics*, 34(17):i757–i765, 2018.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

F Ann Ran, Le Cong, Winston X Yan, David A Scott, Jonathan S Gootenberg, Andrea J Kriz, Bernd Zetsche, Ophir Shalem, Xuebing Wu, Kira S Makarova, et al. In vivo genome editing using staphylococcus aureus cas9. *Nature*, 520(7546):186–191, 2015.

Jeffry D Sander and J Keith Joung. Crispr-cas systems for editing, regulating and targeting genomes. *Nature biotechnology*, 32(4):347, 2014.

Shiraz A Shah, Susanne Erdmann, Francisco JM Mojica, and Roger A Garrett. Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology*, 10(5):891–899, 2013.

Ying Sheng, Pär G Engström, and Boris Lenhard. Mammalian microrna prediction through a support vector machine model of sequence and structure. *PloS one*, 2(9), 2007.

Ritambhara Singh, Cem Kuscu, Aaron Quinlan, Yanjun Qi, and Mazhar Adli. Cas9-chromatin binding information enables more accurate crispr off-target prediction. *Nucleic acids research*, 43(18):e118–e118, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Manuel Stemmer, Thomas Thumberger, Maria del Sol Keyer, Joachim Wittbrodt, and Juan L Mateo. Cctop: an intuitive, flexible and reliable crispr/cas9 target prediction tool. *PloS one*, 10(4), 2015.

Shengdar Q Tsai, Zongli Zheng, Nhu T Nguyen, Matthew Liebers, Ved V Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A John Iafrate, Long P Le, et al. Guide-seq enables genome-wide profiling of off-target cleavage by crispr-cas nucleases. *Nature biotechnology*, 33(2):187, 2015.

Xuebing Wu, David A Scott, Andrea J Kriz, Anthony C Chiu, Patrick D Hsu, Daniel B Dadon, Albert W Cheng, Alexandro E Trevino, Silvana Konermann, Sidi Chen, et al. Genome-wide binding of the crispr endonuclease cas9 in mammalian cells. *Nature biotechnology*, 32(7):670, 2014.

Xiaojun Xu, Dongsheng Duan, and Shi-Jie Chen. Crispr-cas9 cleavage efficiency correlates strongly with target-sgrna folding stability: from physical mechanism to off-target assessment. *Scientific reports*, 7(1):1–9, 2017.

Xiao-Hui Zhang, Louis Y Tee, Xiao-Gang Wang, Qun-Shan Huang, and Shi-Hua Yang. Off-target effects in crispr/cas9-mediated genome engineering. *Molecular Therapy-Nucleic Acids*, 4:e264, 2015.

A APPENDIX

In Figures 4 and 5, we illustrate how CNNs can be also adapted to genome editing by considering each sgRNA-DNA sequence as a black and white image since sgRNA-DNA sequences can be stored in a matrix Lin & Wong (2018). Figure 4 shows the encoding process of the sgRNA-DNA. Figure 5 illustrates a one-hot encoded matrix containing both sgRNA and DNA information for a sequence of 23 nucleobases with the last 3 nucleobases being the PAM.

Figures 6 and 7 present the Receiver Operating Characteristic (ROC) curves of all machine learning models being compared for both considered encodings, 4×23 and 8×23 for the experiments on the CRISPOR data set. The ROC Area Under the Curve (AUC) value of each model is also presented within each figure. For all tested neural networks and the Random Forest (RF) model, the encoding 8×23 leads to greater AUC values than the 4×23 encoding. The biggest difference occurs for the CNN model of Lin and Wong. Lin & Wong (2018), where the proposed 8×23 encoding leads to a 12% increase of the AUC statistic.

Figures 8 and 9 present the Receiver Operating Characteristic (ROC) curves of all machine learning models being compared for both considered encodings, 4×23 and 8×23 for the transfer learning experiments on the GUIDE-seq data set. The ROC Area Under the Curve (AUC) value of each model is also presented within each figure. The difference in the results provided by the two competing encodings was greater in this experiment with an improvement of the AUC statistic value up to 35% (for the FNN with 5 layers). The presented curves show that all the models gain in accuracy when the 8×23 encoding is used. The model of Lin and Wong Lin & Wong (2018), inherited from the authors GitHub page, did not perform as well as the other models. We however did not want to change the authors' implementation as it was supposed to be the exact model described in their experiments Lin & Wong (2018), which is considered here for state of the art comparison.

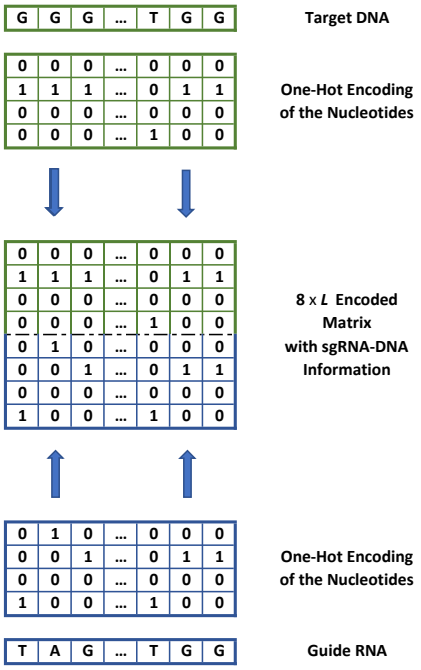


Figure 4: An example of encoding a sgRNA-DNA sequence pair of length L into a $8 \times L$ matrix. Each nucleotide of both sgRNA and DNA sequences is one hot-encoded, giving us two sub-matrices of size $4 \times L$. The two sub-matrices are merged to form the final matrix of size $8 \times L$ containing sgRNA-DNA information. This matrix is then used as input for the FNN and CNN modeling.

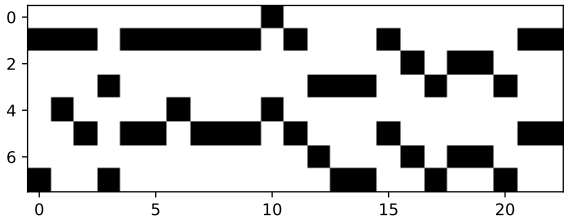


Figure 5: An example of visualization of a sgRNA-DNA sequence pair encoded into a 8×23 matrix. This matrix can be used to assess the differences between sgRNA and DNA sequences during the neural network off-target prediction.

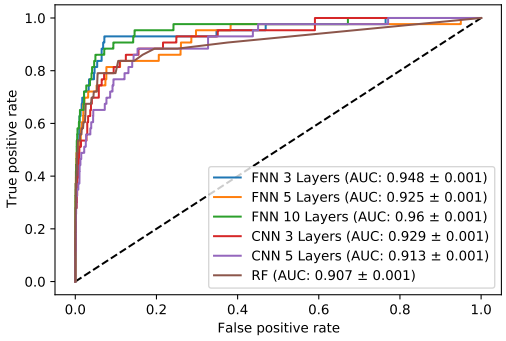


Figure 6: ROC curves for the FNN models with 3, 5 and 10 layers, the CNN models with 3 and 5 layers, the CNN model of Lin and Wong and the RF model obtained on the CRISPOR data set using the 4×23 data encoding. The AUC ROC value for each model is shown within the figure.

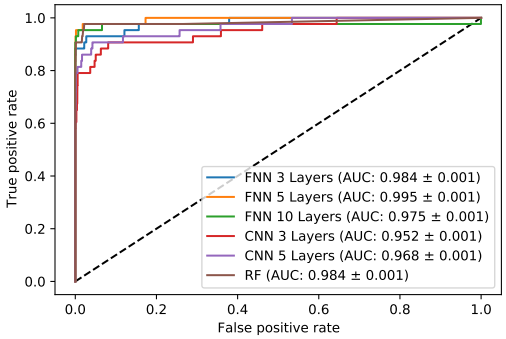


Figure 7: ROC curves for the FNN models with 3, 5 and 10 layers, the CNN models with 3 and 5 layers, the CNN model of Lin and Wong and the RF model obtained on the CRISPOR data set using the 8×23 data encoding. The AUC ROC value for each model is shown within the figure. The 8×23 encoding always provided greater values of the AUC ROC statistic than the 4×23 encoding, see Figure 6.

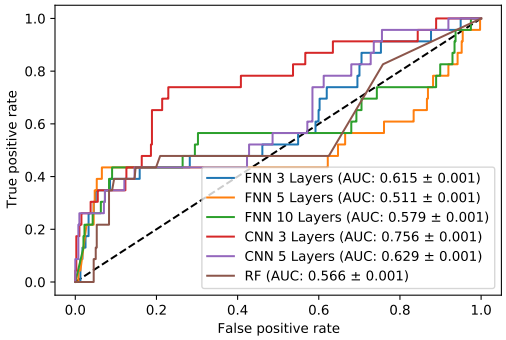


Figure 8: ROC curves for the FNN models with 3, 5 and 10 layers, the CNN models with 3 and 5 layers, the CNN model of Lin and Wong and the RF model obtained on the GUIDE-seq data set using the 4×23 data encoding. The AUC ROC value for each model is shown within the figure.

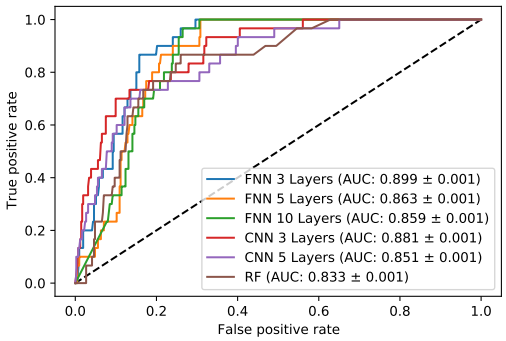


Figure 9: ROC curves for the FNN models with 3, 5 and 10 layers, the CNN models with 3 and 5 layers, the CNN model of Lin and Wong and the RF model obtained on the GUIDE-seq data set using the 8×23 data encoding. The AUC ROC value for each model is shown within the figure. The 8×23 encoding always provided greater values of the AUC ROC statistic than the 4×23 encoding, see Figure 8.