Final Project Report for CS 184A/284A, Fall 2021

Project Title: Predicting Off-target activity scores in CRISPR-Cas9 gRNA gene editing
Project Number: #18

Student Name(s)
Julius Aguma, 38208345, jaguma@uci.edu

1. Introduction and Problem Statement:
Motivation

The clustered regularly interspaced short palindromic repeats (CRISPR) technology has revolutionized gene editing and brought new hope for future disease prevention, aging, and even sparked talk of 'designer dna' (the trendy name for potential aesthetic gene editing).

Perhaps the biggest flaw in the Crispr technology is the inability to accurately predict the effects of editing a certain gene using a specific guide Ribonucleic acid(RNA). Many solutions to the problem(predicting off-target activity) have been proposed. Included among these are the Elevation and Crispor algorithms that use machine learning to predict off-target activity for end-to-end gRNA design.

Problem Statement

Given a DNA target-gRNA pair, can we predict the risk of off-target activity if the Crispr technology were to be deployed for gene editing using that specific pair. This project looks to apply different artificial neural network models towards predicting off-target activity when using Crispr gRNAs. We use GUIDE-seq datasets[1] to train and test different models then evaluate their performance against each other and against the Elevation and Crispor algorithms.

2. Related Work:

As mentioned above, two very relevant works are Elevation and Crispor algorithms introduced in papers [1] and [2] respectively. [1] breaks the problem of predicting off-target activity into three steps namely; (1) given a specific genome, search and filter the entire genome for potential gRNA targets. That is, finding possible locations for a gene edit, (2) Score each potential target with the gRNA in question for possible off-target activity. This involves finding mismatches and calculating the likelihood for off-target activity after the edit. This is closely related to what we will be doing in this project, (3) And lastly we aggregate the scores of all potential targets with that one gRNA into one gRNA score that tells us how effective the gRNA will be for the particular genome. All three steps are summarized in figure 1 below.

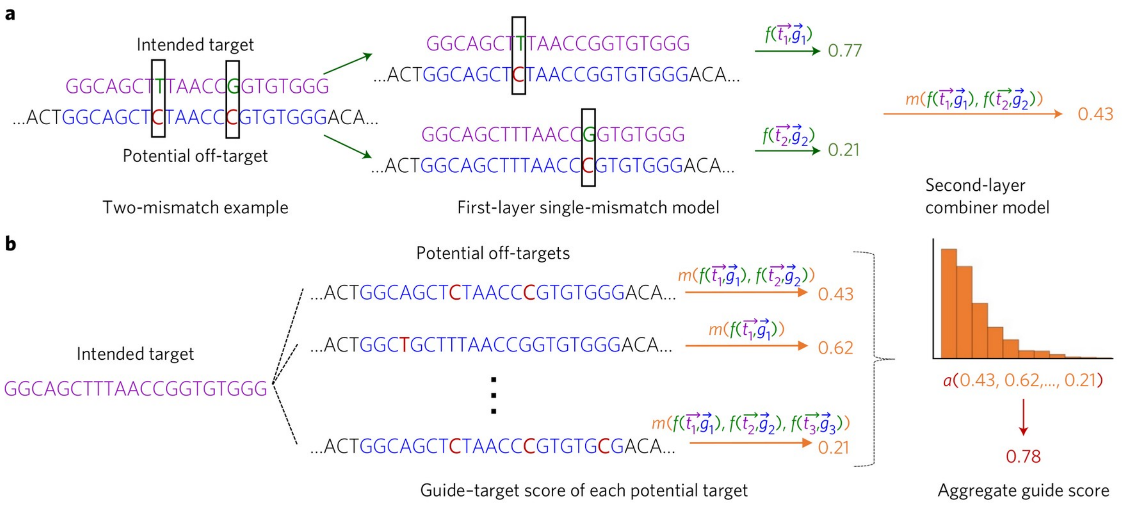

a
Intended target
GGCAGCTTTAACCGGTGTGGG
...ACTGGCAGCTCTAACCCGTGTGGGACA...
Potential off-target
Two-mismatch example
GGCAGCTTTAACCGGTGTGGG
...ACTGGCAGCTCTAACCGGTGTGGGACA...
GGCAGCTTTAACCGGTGTGGG
...ACTGGCAGCTTTAACCCGTGTGGGACA...
First-layer single-mismatch model
$f(\vec{t_1},\vec{g_1})$ 0.77
$f(\vec{t_2},\vec{g_2})$ 0.21
$m(f(\vec{t_1},\vec{g_1}), f(\vec{t_2},\vec{g_2}))$ 0.43
Second-layer combiner model
b
Intended target
GGCAGCTTTAACCGGTGTGGG
Potential off-targets
...ACTGGCAGCTCTAACCCGTGTGGGACA...
$m(f(\vec{t_1},\vec{g_1}), f(\vec{t_2},\vec{g_2}))$ 0.43
...ACTGGCTGCTTTAACCGGTGTGGGACA...
$m(f(\vec{t_1},\vec{g_1}))$ 0.62
...ACTGGCAGCTCTAACCCGTGTGCGACA...
$m(f(\vec{t_1},\vec{g_1}), f(\vec{t_2},\vec{g_2}), f(\vec{t_3},\vec{g_3}))$ 0.21
Guide–target score of each potential target
$a(0.43, 0.62,..., 0.21)$
0.78
Aggregate guide score

Figure 1: A summary of the elevation model functionality. This project is most closely related to part a.

[2] is much like a survey but with a web application as an output. The paper compares a number on algorithms for the problem of predicting off-target activity and as a result, they develop an end-to-end tool, CRISPOR (http://crispor.org), for this task. A summary of the comparison is shown in figure 2.

One particular algorithm, the CFD score, reviewed in [2] is of great significance to this project because it is deployed as a benchmark score for each target-gRNA score as is detailed in the data set section to follow.

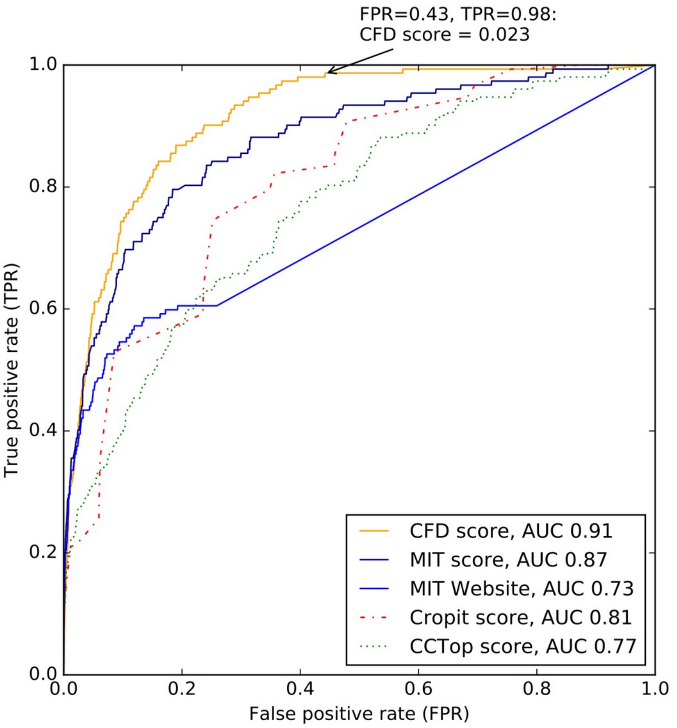
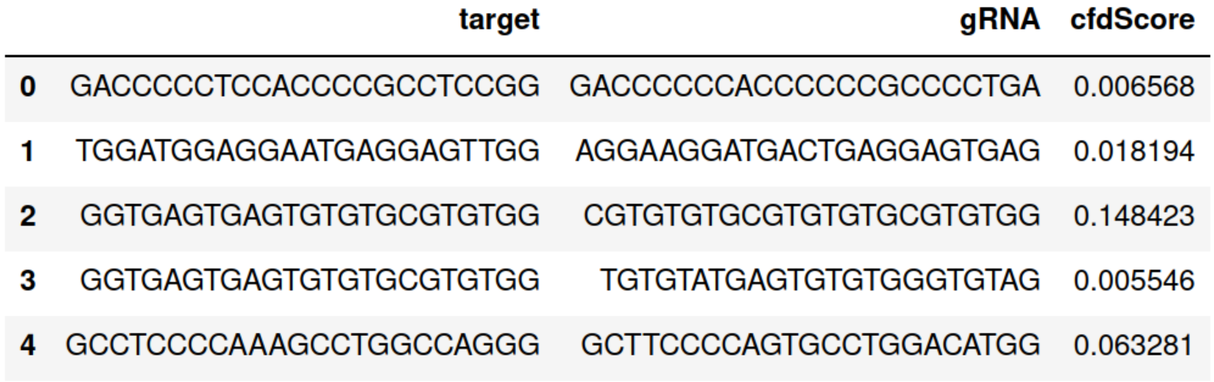


Figure 2: A comparison of the state-of-the-art algorithms for predicting off-target activity. The CFD score algorithm is the same benchmark this project uses to predict off-target activity.

3. Data Sets

We use data generated by [1] for the task of predicting off-target activity. [1] generated this data from the GUIDE-seq analysis tool that is publicly available and is one of the most popular dataset in genome study. Figure 3 is a table of the data we use for our algorithm with;



	target	gRNA	cfdScore
0	GACCCCCTCCACCCCGCCTCCGG	GACCCCCCACCCCCCGCCCCTGA	0.006568
1	TGGATGGAGGAATGAGGAGTTGG	AGGAAGGATGACTGAGGAGTGAG	0.018194
2	GGTGAGTGAGTGTGTGCGTGTGG	CGTGTGTGCGTGTGTGCGTGTGG	0.148423
3	GGTGAGTGAGTGTGTGCGTGTGG	TGTGTATGAGTGTGTGGGTGTAG	0.005546
4	GCCTCCCCAAAGCCTGGCCAGGG	GCTTCCCCAGTGCCTGGACATGG	0.063281

Figure 3: 26052 Target-gRNA pairs with corresponding CFD scores. This data was generated by [1] andis available here: https://github.com/microsoft/Elevation/tree/master/CRISPR/data/offtarget/Haeussler

- Column 1: 23-long potential genome targets primed for edit with the Cas9/Crispr tool

- Column 2: Corresponding gRNAs for each potential target

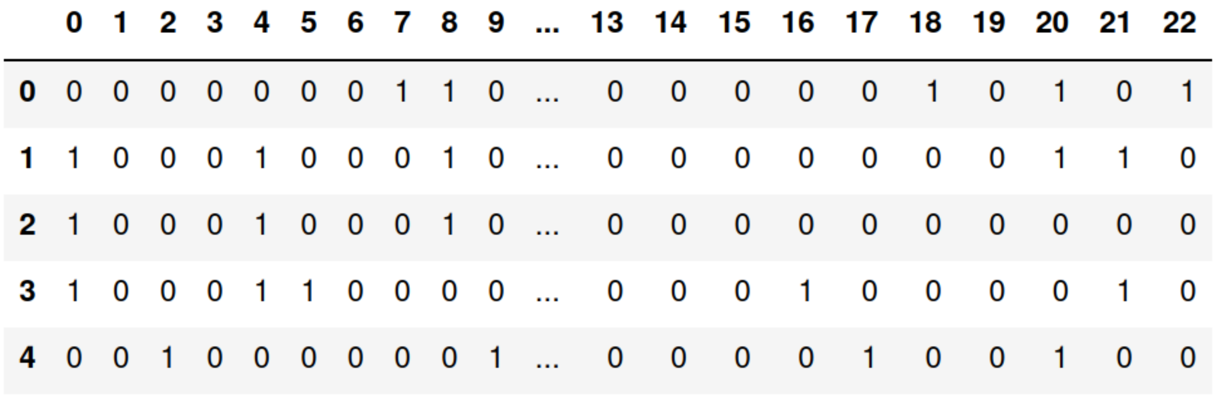
- Column 3: CFD scores as scored by the Cutting Frequency Determination algorithm[3].

The next section details the preprocessing steps that converted this raw data into model input.

4. Description of Technical Approach

Preprocessing:

The target-gRNA pairs are processed by a one-hot encoder where each mismatch is a 1 and every match in the pair is a 0. Figure 4 shows an example for the first 5 pairs shown in figure 3.



	0	1	2	3	4	5	6	7	8	9	...	13	14	15	16	17	18	19	20	21	22
0	0	0	0	0	0	0	0	1	1	0	...	0	0	0	0	0	1	0	1	0	1
1	1	0	0	0	1	0	0	0	1	0	...	0	0	0	0	0	0	0	1	1	0
2	1	0	0	0	1	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	1	1	0	0	0	0	...	0	0	0	1	0	0	0	0	1	0
4	0	0	1	0	0	0	0	0	0	1	...	0	0	0	0	1	0	0	1	0	0

*Figure 4: Examples of the one-hot encoding that converts a target-gRNA pair into a 1*23 array of 1s (for mismatch) and 0s (match).*

After the encoding the data is split into training and testing data at 80% of the pairs for training and 20% for testing.

Linear Regression Models:

The project compares the performance of three linear regression models on scoring the target-gRNA pairs data, that is, giving a likelihood score for off-target activity for each pair. The three models are summarized below.

1. FeedForward Neural Network. The first is a simple FFN model with three layers consisting of 16, 8, 1 neurons in the respective layers.

2. Multi-Layer Perceptron. The second model is a simple MLP with the same architecture as the FFN model, that is, 3 layers(16,8,1 neurons).

3. Recurrent Neural Network. The last of the models is a simple RNN with two layers(16,8).

All three models use the same hyper-parameters(learning rate, number of hidden layers, epochs, batch size, and activation function) and produced a single risk of off-target activity score.

4. Software

- Keras and Tensorflow: Popular open-source python interface that provide pre-written implementations of all three artificial neural network models with keras as the front-end and Tensorflow as the backend.

- Data and Evaluation Visualization tools: panda library, sklearn, and matplot. All popular open-source python libraries

- The following websites contributed implementation code to this project

FFN:	https://vitalflux.com/keras-neural-

	network-for-regression-problem/
MLP:	https://valueml.com/multi-layer-perceptron-by-keras-with-example/
RNN:	https://www.tensorflow.org/guide/keras/rnn

5. Experiments and Evaluation

The project compares three artificial neural network models for the task of predicting off-target activity. For a fair comparison, we hold the hyper-parameters constant through all three models first. We then test the models on two metrics, MSE and MAE. These are summarized below.

Mean Absolute Error:

This is the simple difference between the true CFD score/label and the predicted scores. The metric takes an average of the errors in prediction from every target-gRNA pair in the dataset. Sklearn provides the neat implementation of this metric that we employed for this project.

Mean Square Error:

This is an average of the square of the difference between the true CFD score and the predicted score. Sklearn also provides the neat implementation of this metric that we adopted.

The results from both metrics at a learning rate of 0.001, 35 epochs, 200 batch size, using RELU activation functions on all layers follow.

a. Feed Forward Network

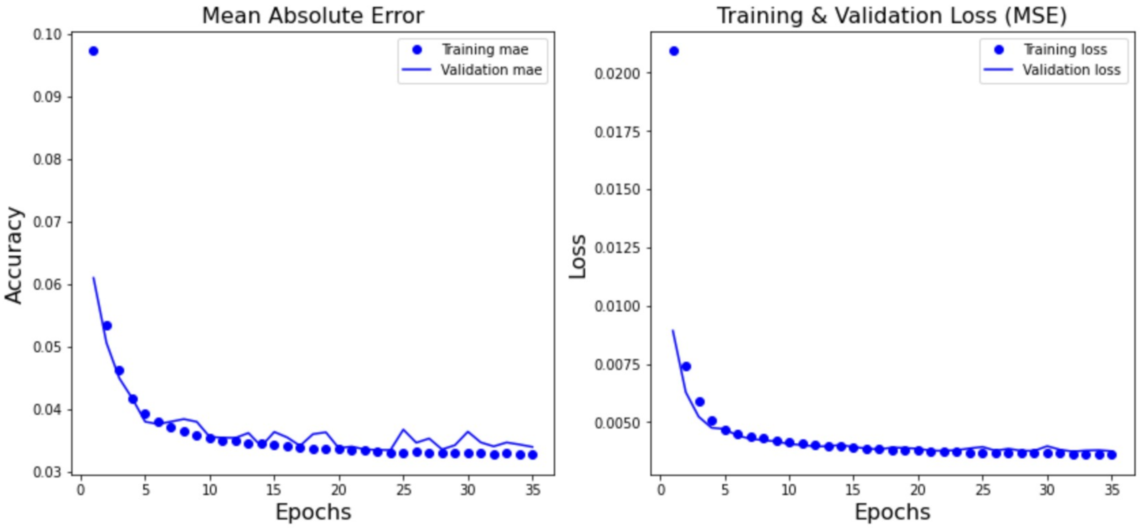


Figure 5: Plots of the performance of the feed forward network with a learning rate of 0.001, and the RELU activation function and a batch size of 200.

b. Multi-layer Perceptron

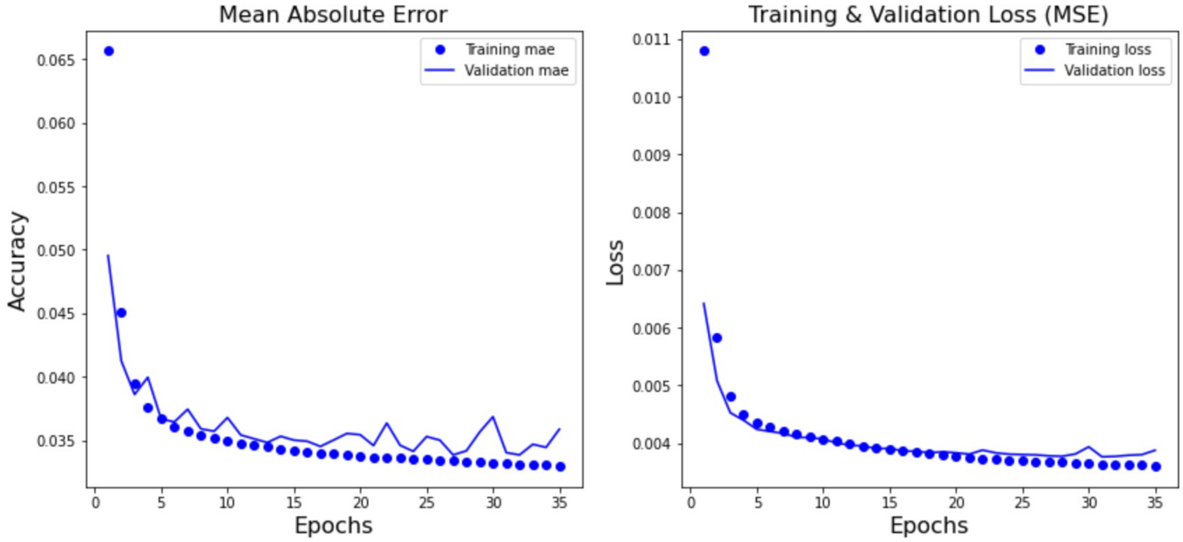


Figure 6: Figure 6: Plots of the performance of the Multi-layer Perceptronnetwork with a learning rate of 0.001, and the RELU activation function and a batch size of 200.

c. Recurrent Neural Network

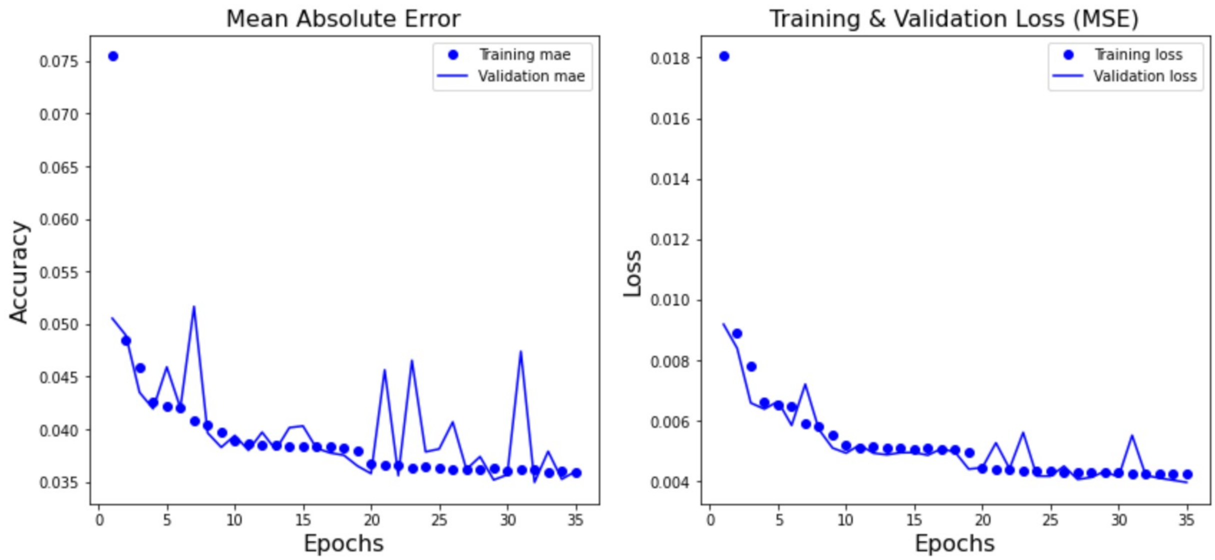


Figure 7: lots of the performance of the recurrent neural network with a learning rate of 0.001, and the RELU activation function and a batch size of 200.

6. Discussion and Conclusion

From the results, while all three models do quite well, we observe that the feed forward network gave us the most stable performance on both training and testing data while the multi-layer perceptron was not far off either. The recurrent network was far more unstable with testing data and took a bit longer to converge even on training data. Perhaps the fixed hyper-parameters are constraining for the RNN. We however did not witness better performance by the RNN anywhere else in the hyper-parameters landscape but maybe future work could deploy better parameter tuning methods to optimize each model.

These results concur with findings from [1] because they too, recommend using a feed forward network for the task of scoring target-gRNA pairs. We however did not get to compare the three models against the elevation-score models because we did not have enough time to implement the spearman weight metric used by [1]. Future work could look into this extension and comparison with other state-of-the-art models.

7. A separate page on *Individual Contributions*

All parts of the project were done solely by Julius Ceasar Aguma.

8. References

Jennifer Listgarten*, Michael Weinstein*, Benjamin P. Kleinstiver, Alexander A. Sousa, J. Keith Joung, Jake Crawford, Kevin Gao, Luong Hoang, Melih Elibol, John G. Doench*, Nicolo Fusi*. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering* Jan 2018, doi:10.1038/s41551-017-0178-6. (*equal contributions, corresponding author)

Jean-Paul Concordet, Maximilian Haeussler, CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens, *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W242–W245, https://doi.org/10.1093/nar/gky354

Doench, John G et al. “Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9.” *Nature biotechnology* vol. 34,2 (2016): 184-191. doi:10.1038/nbt.3437