Q1. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity
Recently in 2015, Cpf1, a single-RNA-guided endonuclease of a class 2 CRISPR-Cas system, was discovered. Compared to Cas9, Cpf1 has several advantages when used in genome editing: it's smaller and simpler, requires only one RNA molecule and creates sticky ends. However, there is only limited knowledge about the relationship between target sequence composition and Cpf1 activity. Prior works focus on predicting Cpf1 activity based on machine learning model trained on a medium-scale dataset, leaving a lot of space for optimization. This paper proposes two deep-learning algorithms, Seq-deepCpf1 and DeepCpf1, to predict the activity of AsCpf1 guide RNAs.

In general, the proposed deep-learning approach outperforms previous machine learning algorithms, on both the authors and published data sets. The Seq-deepCpf1 algorithm is a deep-learning based regression model to predict AsCpf1 activity based on target sequence composition. As the size of training data increases, Seq-deepCpf1 achieves a Spearman correlation coefficient of 0.75 compared to experimentally obtained indel frequencies. Cross-validation results show that, compared to the state-of-the-art approaches (L1/L2/L1L2-regularized linear regression) for Cas9 activity prediction, the Spearman correlation of Seq-deepCpf1 is significant higher. The authors also combine chromatin accessibility with target sequence composition, leading to the design of DeepCpf1. Evaluation results using HEK-plasmid and HCT-plasmid show that DeepCpf1 has substantially better performance than other approaches (with Spearman correlations of 0.87 and 0.77, and AUCs of 0.89 and 0.91). This paper demonstrates that deep-learning helps to improve prediction of CRISPR–Cpf1 guide RNA activity.

I think the limitations of the paper is twofold: the proposed approach could not generalize well on different datasets and requires much manual tuning; and further, the authors provide a comparison between Cas9 and Cpf1 activity prediction. It's kind of indirect since the problem, the training dataset and the model are different.

This paper is the first to handle Cpf1 activity prediction using deep-learning approaches. It is also well-written and provides lots of experiments as supporting materials. The supplementary information enables the other researchers to validate. However, as in my field of research, it appears to be an incremental work based on "In vivo high-throughput profiling of CRISPR-Cpf1 activity".

Q2. Predicting the Cellular Localization Sites of Proteins

1. In this part, we use *decision tree* to do classification based on *scikit-learn.*

2. In the evaluation part, we do 10-fold cross validation: the dataset is randomly split into 10 equal size subsets. And in each of the 10 iterations, a subset is used as validating data to test the model, and the other subsets are used as training.
3. The performance of classification model is described by confusion matrix. Specifically, we take each class label as positive and all other class labels as negative. The result is list below:

“CYT”	Predicted Yes	Predicted No
Actual Yes	201	262
Actual No	226	795

“NUC”	Predicted Yes	Predicted No
Actual Yes	209	220
Actual No	230	825

“MIT”	Predicted Yes	Predicted No
Actual Yes	113	131
Actual No	247	1093

“ME3”	Predicted Yes	Predicted No
Actual Yes	107	56
Actual No	43	1278

“ME2”	Predicted Yes	Predicted No
Actual Yes	18	33
Actual No	32	1401

“ME1”	Predicted Yes	Predicted No
Actual Yes	30	14
Actual No	15	1425

“EXC”	Predicted Yes	Predicted No
Actual Yes	11	24
Actual No	30	1419

“VAC”	Predicted Yes	Predicted No
Actual Yes	1	29
Actual No	42	1412

“POX”	Predicted Yes	Predicted No
Actual Yes	6	14
Actual No	23	1441

“ERL”	Predicted Yes	Predicted No
Actual Yes	0	5
Actual No	0	1479

4. Run our simple demo python program with: ./decision_tree.py yeast.data k, where k is the number of folds in cross validation.

5. For the dataset of 9 numeric attributes, i would also try random forest, which is a ensembling classification method. The voting of multiple tree model could probably increase the prediction result.