Summer Review 5

DeepCRISPR

Guohui Chuai, Hanhui Ma,Jifang Yan, Ming Chen , Nanfang Hong, Dongyu Xue, Chi Zhou, Chenyu Zhu, Ke Chen, Bin Duan, Feng Gu , Sheng Qu, Deshuang Huang, Jia Wei and Qi Liu

Paper link

Reviewed by : Arshdeep Sekhon

[1]Department of Computer Science, University of Virginia

https://qdata.github.io/deep2Read/

CRISPR

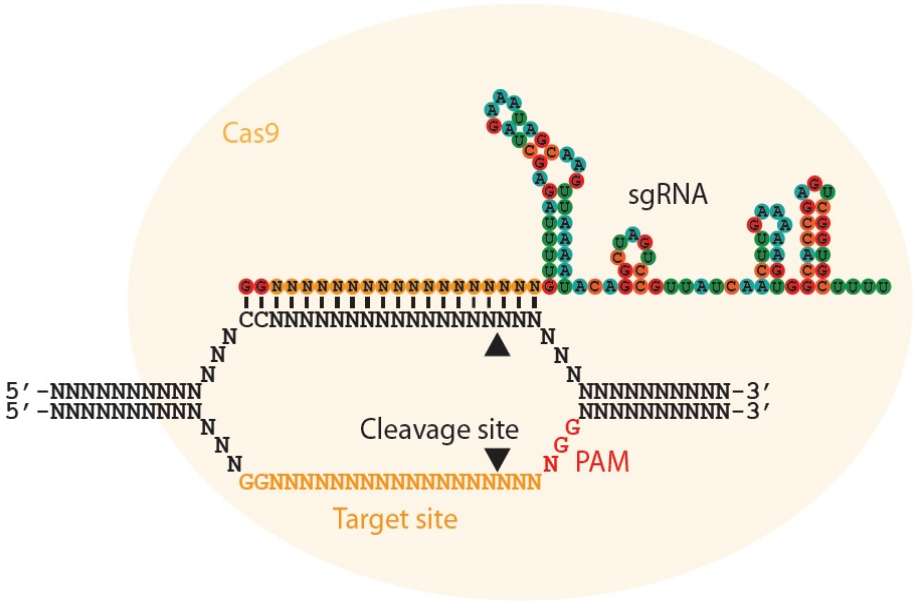


Figure: CRISPR

- sgRNA has 20 nucleotide sequence
- sgRNA attacks matching sequence in the genome + must have a PAM
- can still attack if minor mismatches as well as DNA or RNA bulges: *off targets*

- sgRNA guides Cas9
- optimized design of sgRNA for
  - high specificity (decreasing off target)
  - high sensitivity (increasing on target knockout efficacy)
- DeepCRISPR: unify on-target and off-target site prediction

Related Work: On target knockout efficacy

- Alignment based
- Score based
- Learning based

Related Work: Off target knockout efficacy

- off target scores: CFD Score, MIT Score

CRISPR data issues

- heterogeneous data: different cell types and different data source
- small labeled sample size: few sgRNAs with known knockout efficacies, experimentally expensive to collect
- data imbalance issues: small off target sites in comparison to all sequences
- the leading sequence and epigenetic features: unclear roles

Deep unsupervised learning for sgRNA representation: sgRNA Encoding

- 20-bp sgRNA sequences with an NGG PAM
- These data account for 0.68 billion sgRNA sequences with different epigenetic information curated from 13 human cell types
- DCDNN based autoencoder: unsupervised, pre-trained parent network

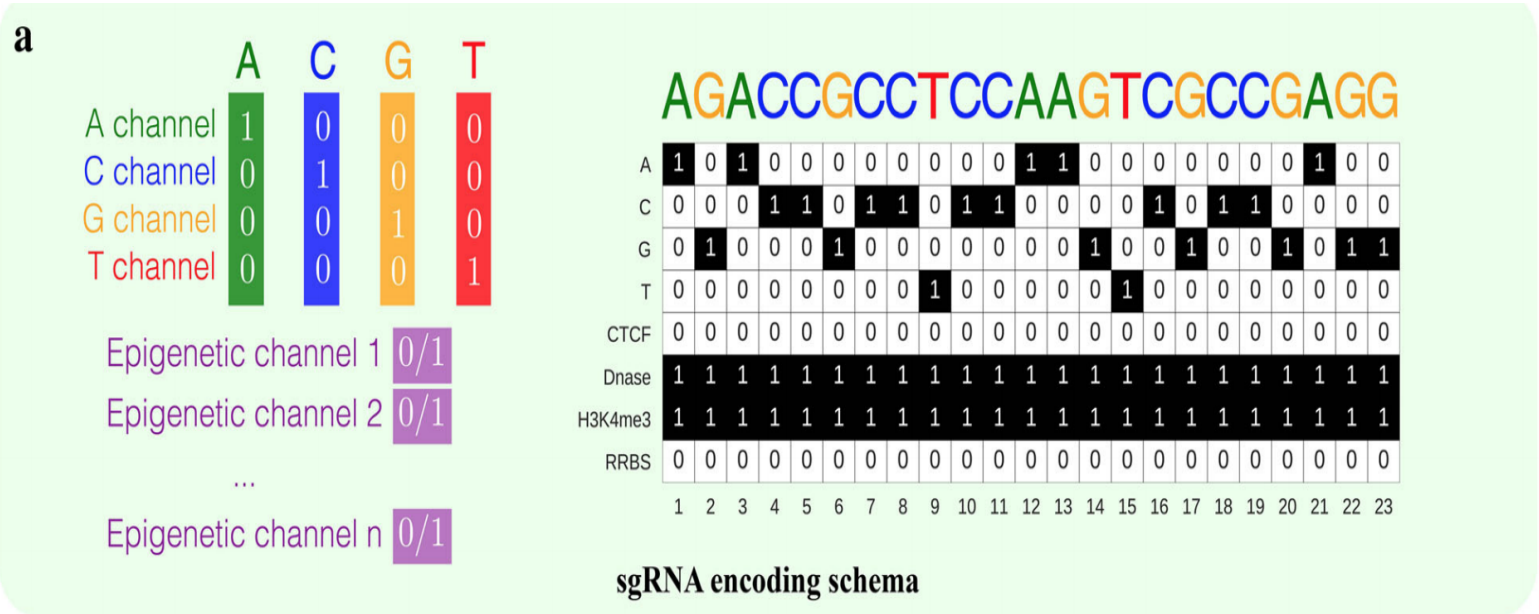
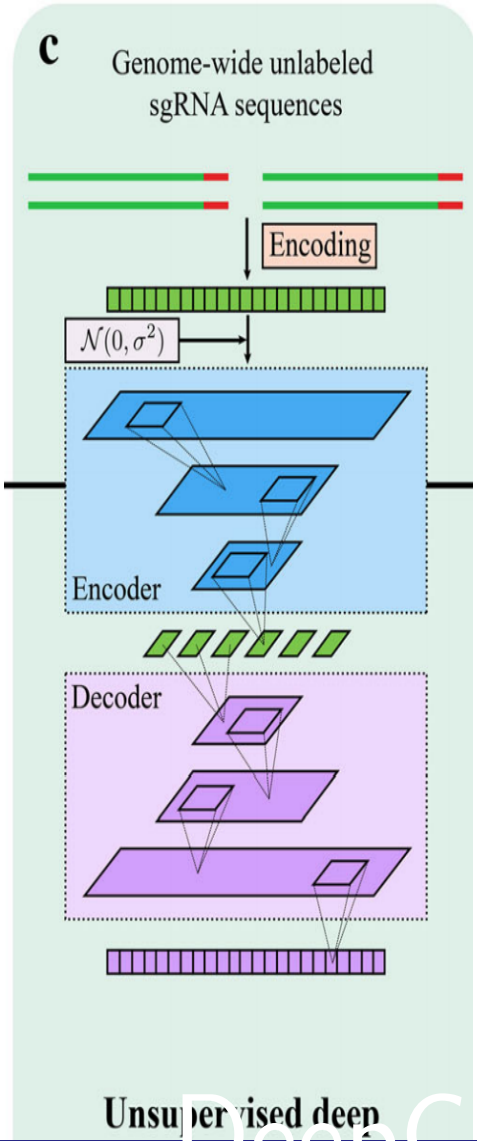


Figure: encoding scheme

Deep unsupervised learning for sgRNA representation: sgRNA Encoding



on target knockout efficacy prediction

hybrid model

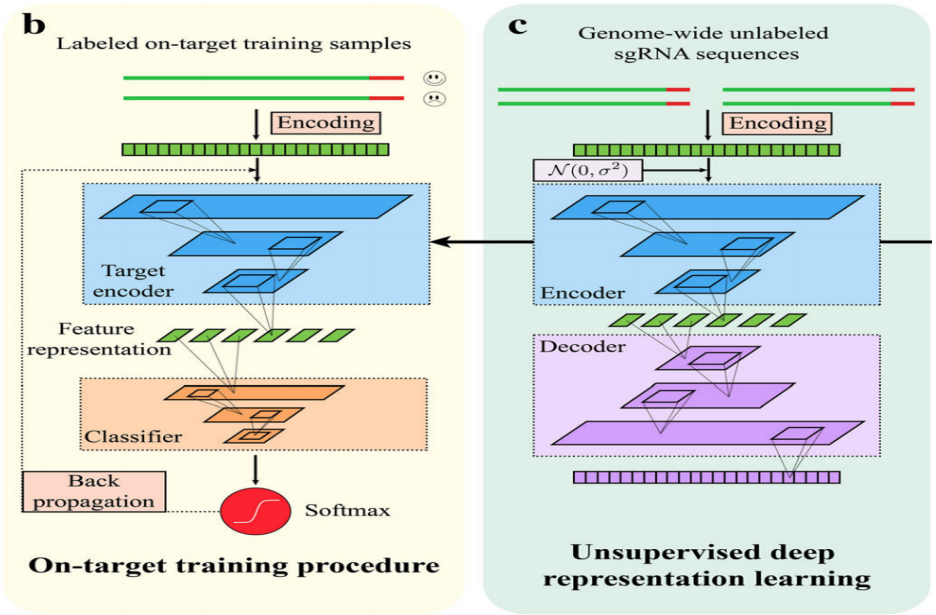


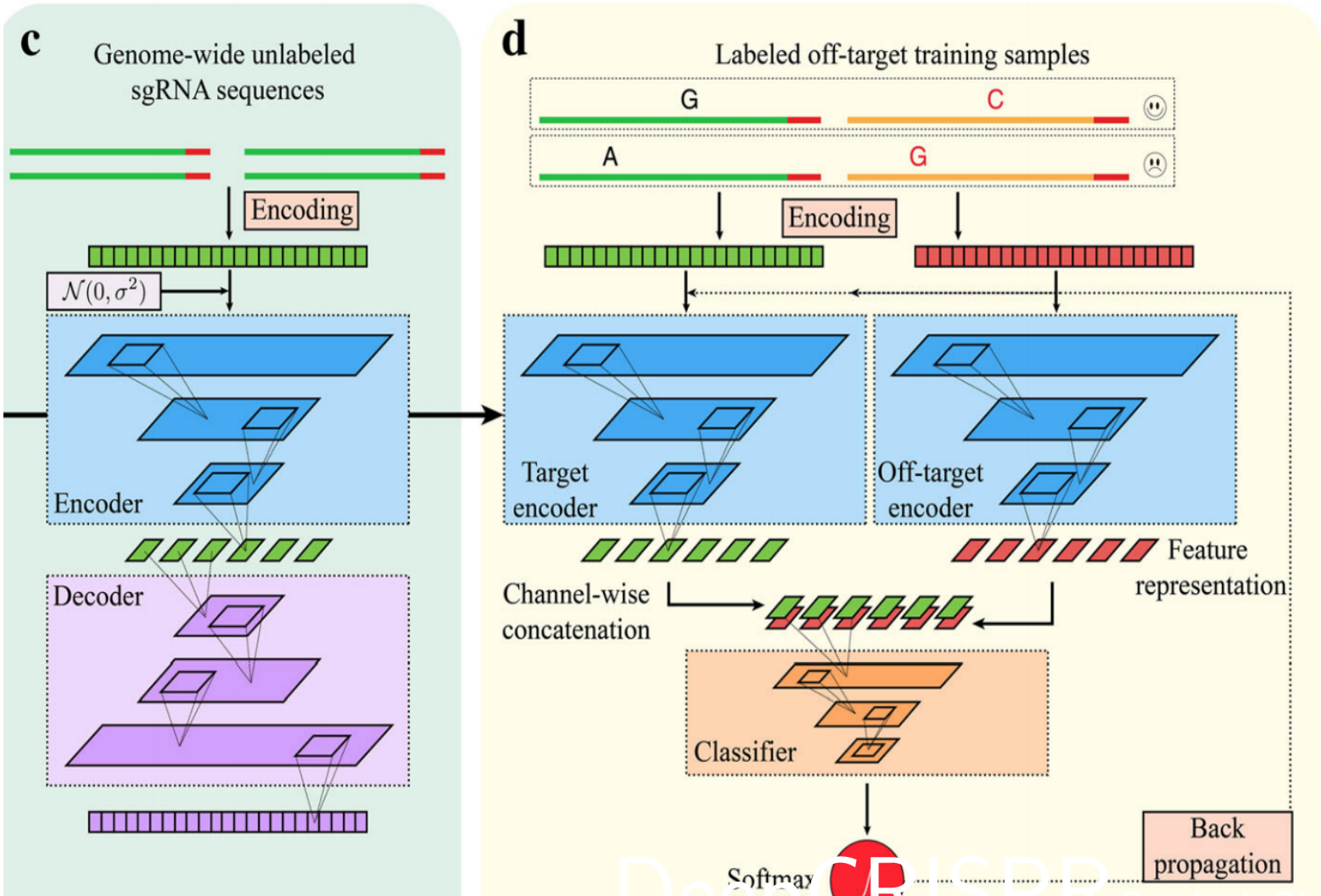
Figure: On target model

the labeled sgRNA dataset contains 0.2 million sgRNAs with known knockout efficacies. This dataset was generated from 15,000 sgRNAs across 1071 genes with known knockout efficacies in a data augmentation manner [1]

[1] Considering that sgRNA with two mismatches in the first two positions from the 5

off target knockout efficacy prediction

- a given sgRNA and its one possible off-target locus as a sample pair
- hybrid model



Results for on target: Testing Scenario 1,2,3

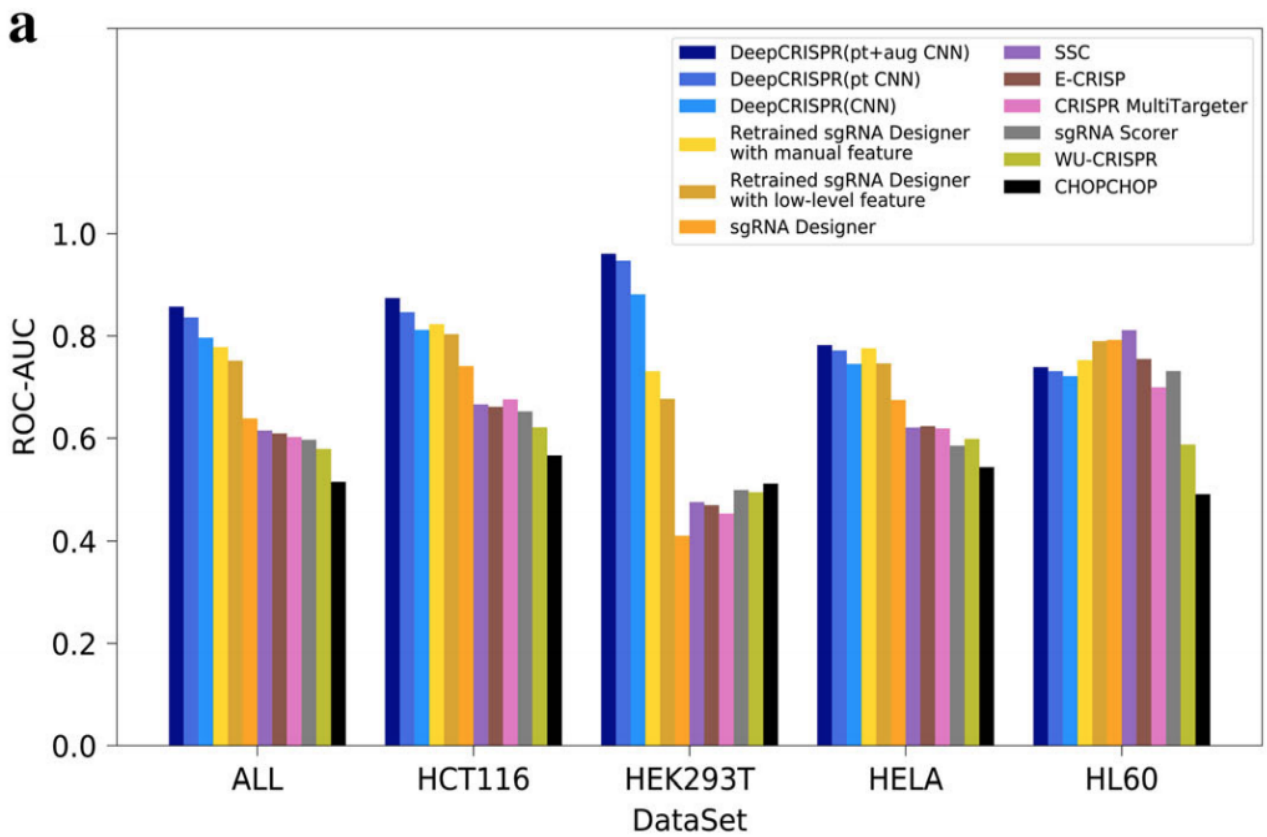


Figure: On target results: with and without pretraining with unsupervised DCCNN

Results for on target: Testing Scenario 4,5

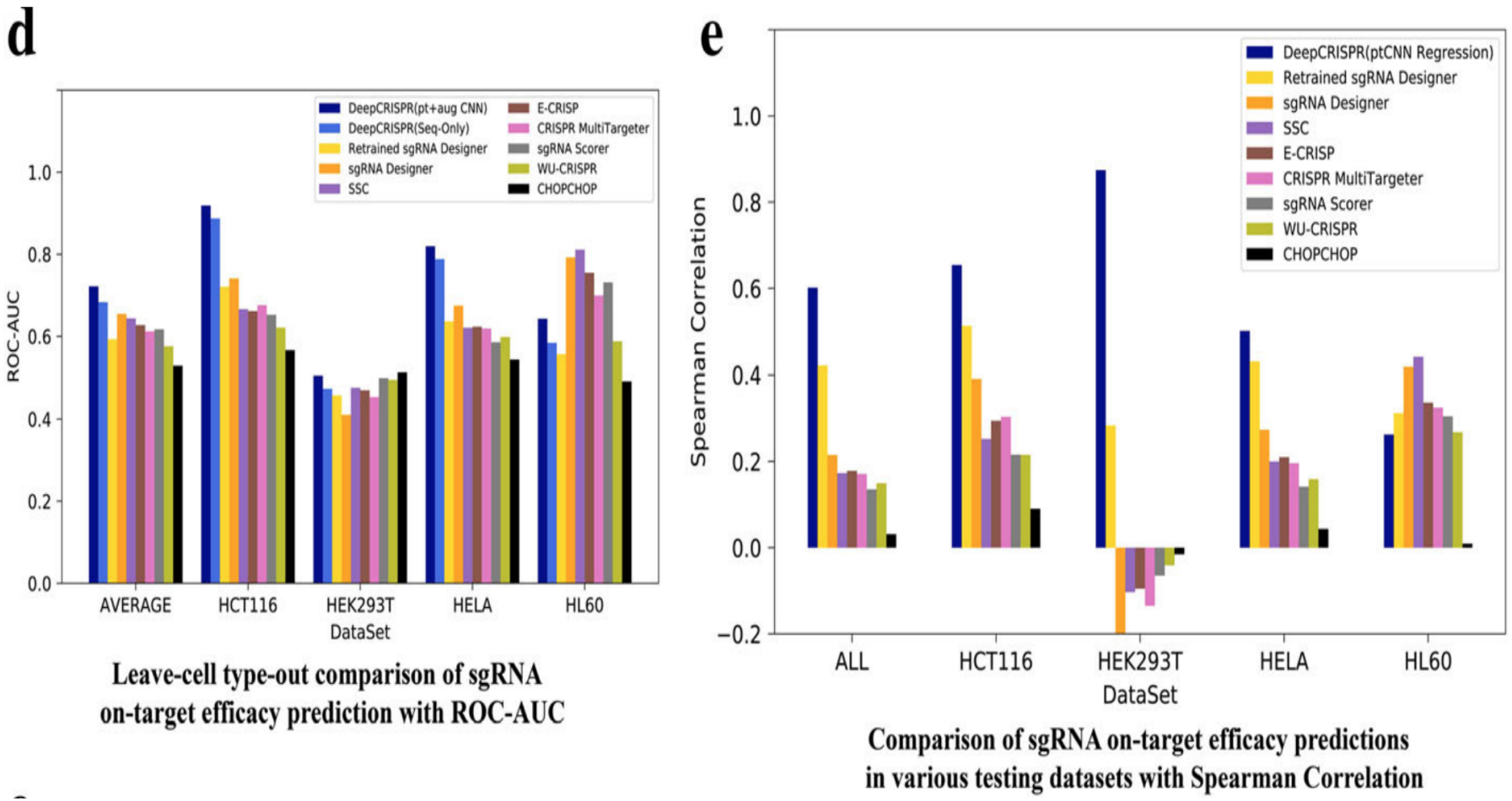


Figure: Leave one cell out, Regression formulation

Results for on target: Testing Scenario 8

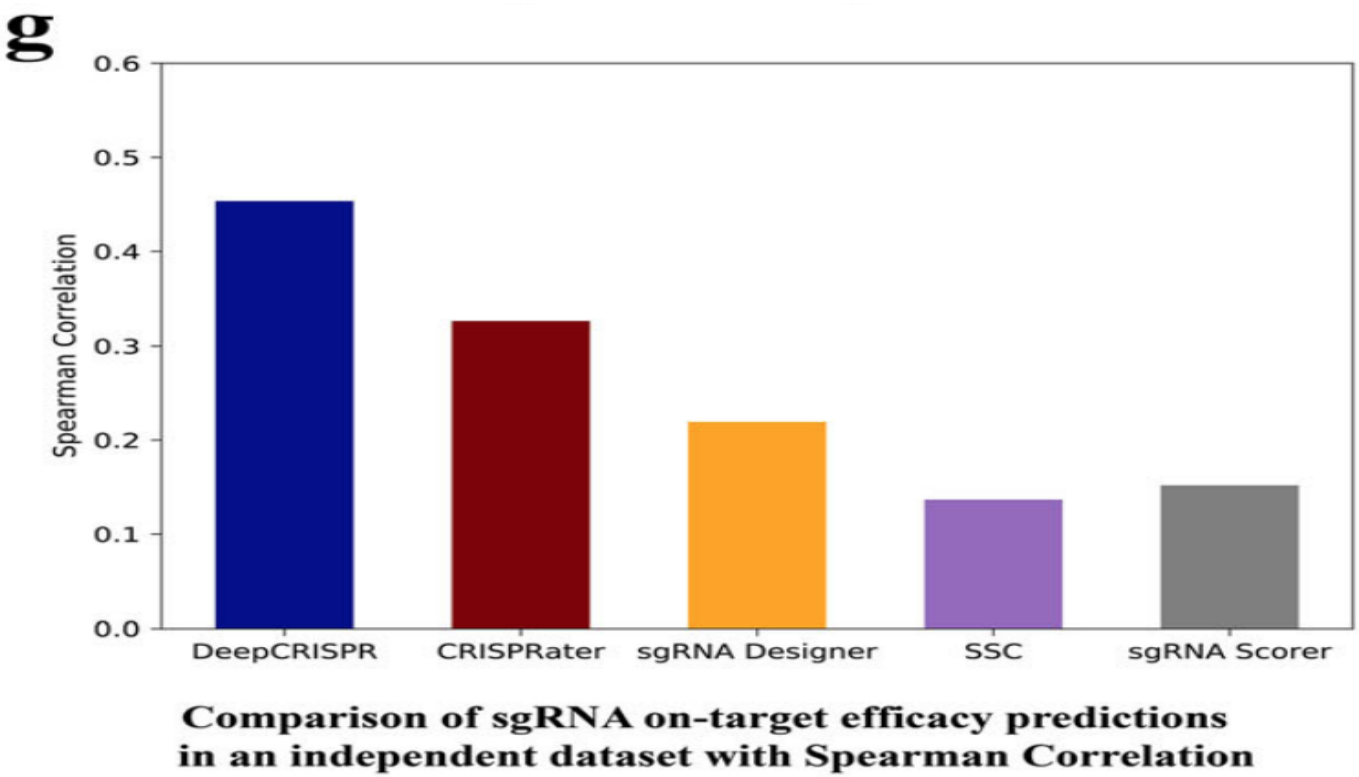
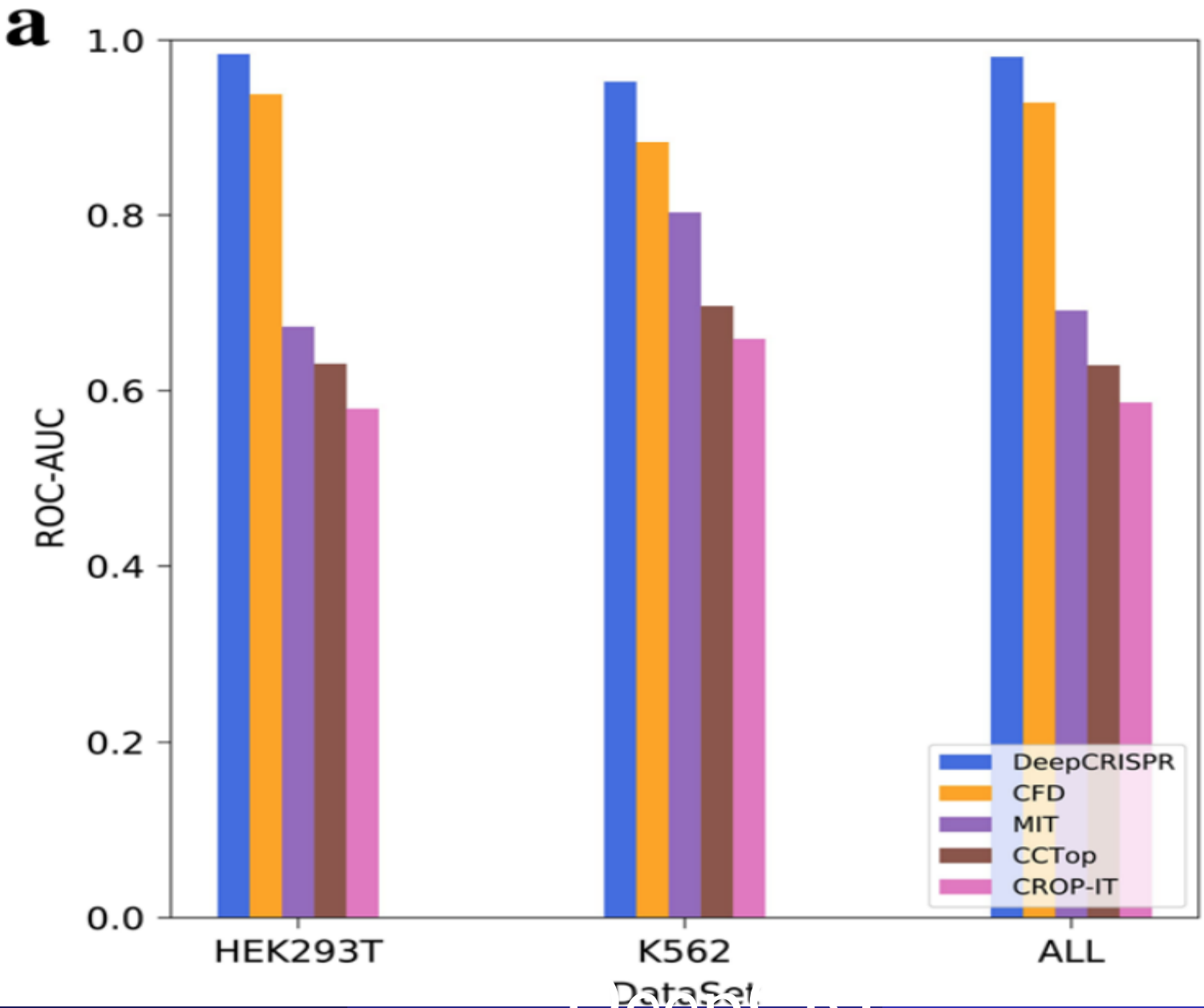


Figure: Independent Dataset: This dataset, reported by utilizing fluorescent reporter knock-out assays with verification at selected endogenous loci for sgRNA knockout efficacy measurement, contains a total of 425 sgRNAs for HEL cells: Both the cell type and data distribution are different and the sgRNAs do not overlap the former datasets.

Off target profile detection

- sgRNA whole-genome off-target profile using GUIDE-seq,Digenome-seq, BLESS, HTGTS, and IDLV
- off-target sites are labeled as “1” and the others are labeled as “0”
- off-target sites are labeled with the targeting efficacies measured with indel frequency detected by different assays
- bootstrapping algorithm for imbalanced dataset
- baselines: MITscore, CFDscore, CROP-IT, CCTop

Testing Scenario 1



Testing Scenario 1

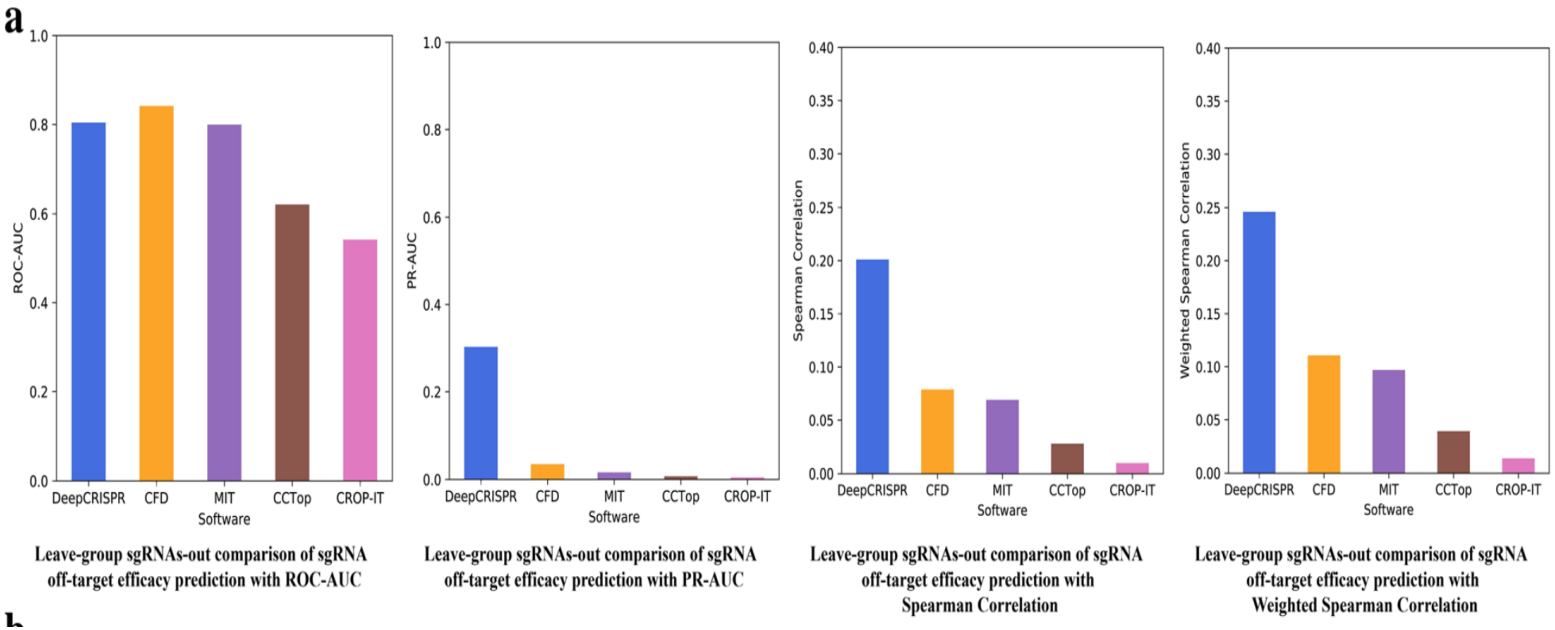


Figure: a Leave sgRNAs group out comparison of sgRNA off-target efficacy prediction with ROC-AUC, PR-AUC, Spearman correlation, and weighted Spearman correlation

Testing Scenario 1

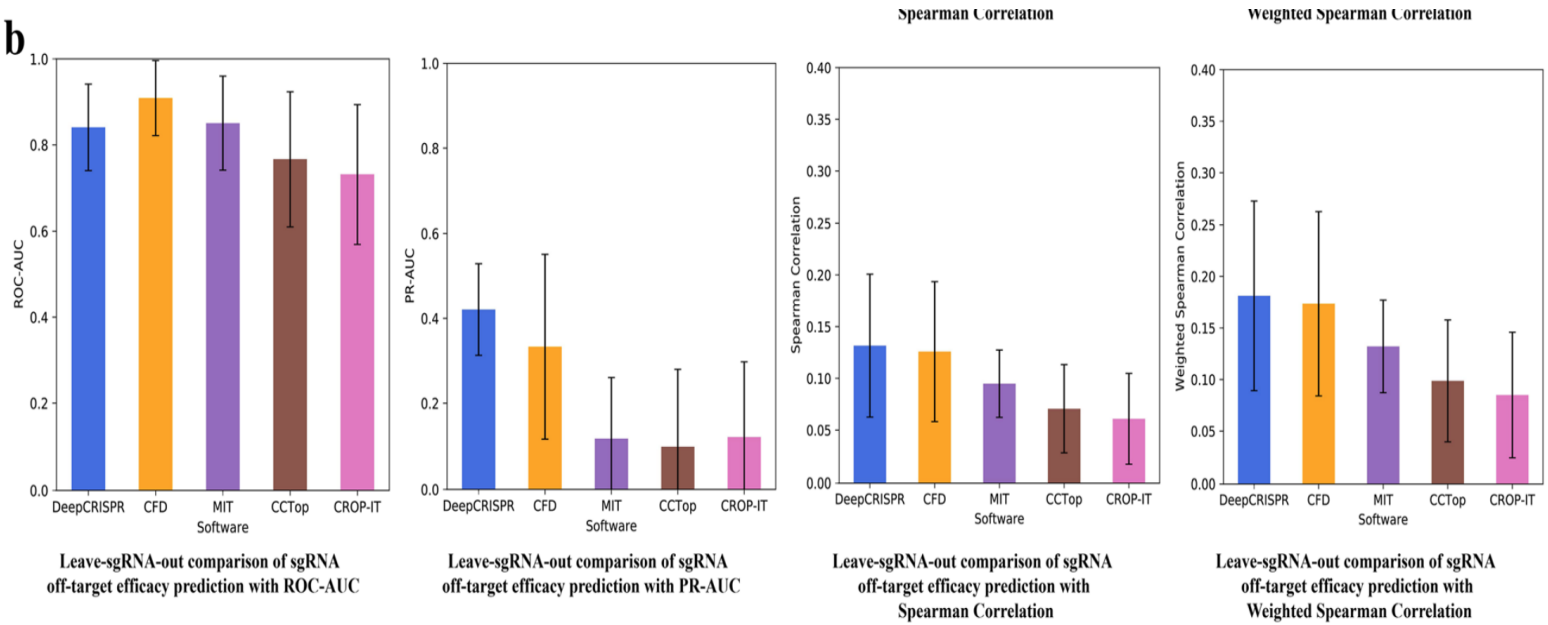


Figure: Leave sgRNA out comparison of sgRNA off-target efficacy prediction with ROC-AUC, PR-AUC, Spearman correlation, and weighted Spearman correlation.

Interpretation

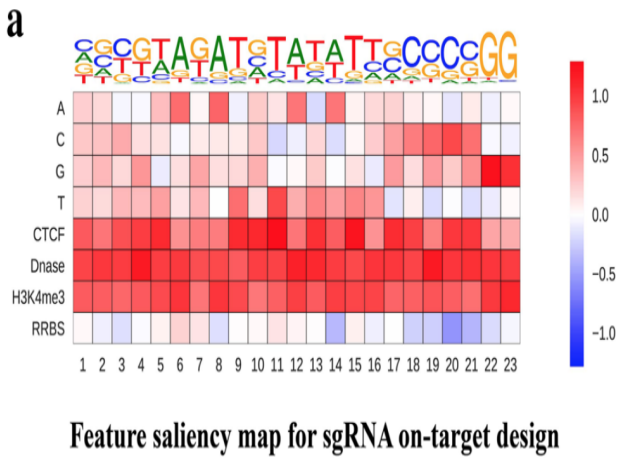
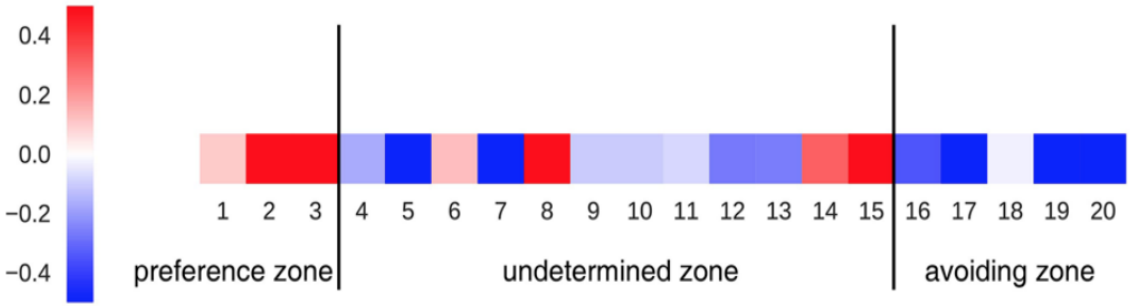


Figure: Saliency map for on target

- it has the same preferences in the variable nucleotides of the PAM NGG for high efficacy sgRNA, where cytosine is favored and thymine is disfavored. This is consistent with existing in vitro and in vivo studies
- Thymines are disfavored at the four positions closest to the PAM, consistent with the fact that multiple uracils in the spacer lead to low sgRNA expression
- It has a general preference for an open chromatin structure, as indicated by the feature saliency map of CTCF, DNase, and H3K4me3

Interpretation



Averaged nucleotide substitution saliency map for sgRNA off-target design

Figure: an averaged nucleotide substitution rate map to indicate their effect on the occurrence of off-target cleavage

- divided this feature map into three nucleotide substitution zones, i.e., off-target preference zone (positions 13), undetermined zone (positions 415), and off-target avoiding zone (positions 1620). Although this map was obtained from limited samples, we observed that the nucleotide mutations occurring near the PAM are prone to avoid off-target sites in a position and nucleotide identity-dependent manner. This is consistent with previous findings that changing the nucleotides far from the PAM usually has little effect on sgRNA efficacy

Interpretation

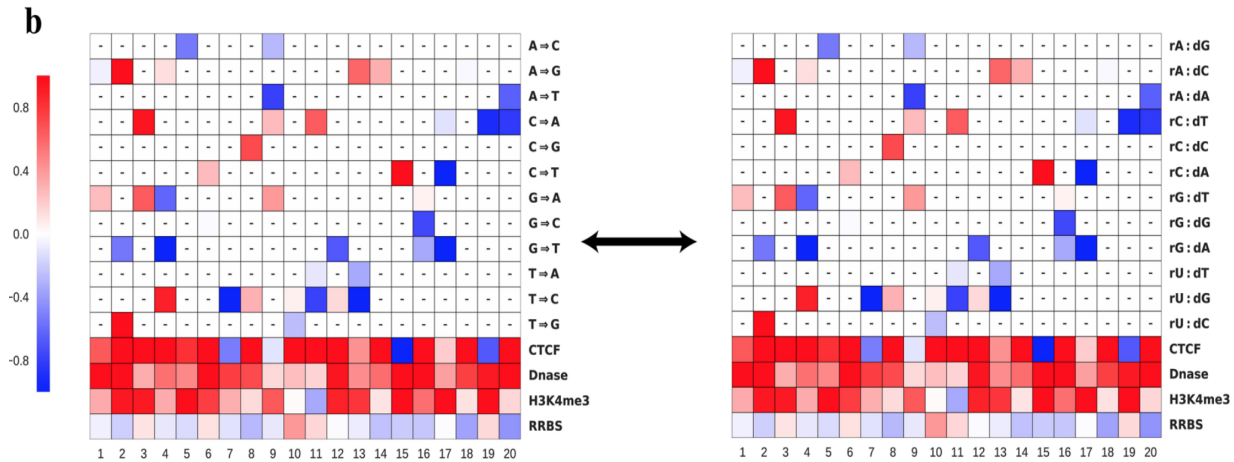


Figure: Saliency map for on target

- DeepCRISPR identified a preference for purine:- purine mismatches to avoid off-target sites with statistical significance, including the substitution G− >C(corresponding to rG:dG in a traditional heatmap, as previously reported) and substitution G− >T (corresponding to rG:dA in a traditional heatmap) at position 16.
- Besides these consistent findings, saliency map identified five nucleotide substitutions preferring off-targets in the off-target preference zone and eight nucleotide substitutions avoiding off-targets in the off-target avoiding zone, including the two nucleotide substitutions G− >C andG− >T at position 16.