Generating, modeling, and evaluating a large-scale set of CRISPR/Cas9 off-target sites with bulges

Ofir Yaish [1] **and Yaron Orenstein** [2,3]✉

[1]School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel, 8410501; [2]Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel, 5290002; [3]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel, 5290002

CRISPR/Cas9 system is widely used in a broad range of gene-editing applications. While the CRISPR editing technique is quite accurate in the target region, there may be many unplanned off-target sites (OTS). Consequently, a plethora of high-throughput experimental assays have been developed to measure OTS in a genome-wide manner. Based on these experimental data, computational methods have been developed to predict OTS given a guide RNA and a reference genome. However, these methods are highly inaccurate when considering OTS with bulges due to limited data compared to OTS without bulges. Recently, CHANGE-seq, a new *in vitro* experimental technique to detect OTS, was used to produce a dataset of unprecedented scale and quality (more than 200,000 OTS over 110 guide RNAs). In addition, the same study included *in cellula* GUIDE-seq experiments for 58 of the guide RNAs. But, while the CHANGE-seq data included more than 20,000 OTS with bulges, the GUIDE-seq data did not include any OTS with bulges. Here, we fill this gap by generating the most comprehensive GUIDE-seq dataset with bulges, and training and evaluating state-of-the-art machine-learning models that consider OTS with bulges. We first reprocessed the publicly available experimental raw data of the CHANGE-seq study to generate 20 new GUIDE-seq datasets, and more than 450 OTS with bulges among the original and new GUIDE-seq experiments. We then trained various machine-learning models, evaluated their performance on multiple datasets, and demonstrated their state-of-the-art performance both *in vitro* and *in cellula*. Last, we visualized the key features learned by our models on OTS with bulges. Our data and models will be instrumental to any future off-target study considering bulges.

CRISPR off-target | GUIDE-seq | CHANGE-seq | Deep neural networks
Correspondence: ***yaron.orenstein@biu.ac.il***

Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 system is the preeminent gene-editing technology in recent years due to its low cost, high efficiency and effectiveness, and ease of use (1; 2; 3; 4; 5). This system is composed of a Cas9 protein and an associated single guide RNA (sgRNA). The sgRNA guides the Cas9 to a target DNA sequence, and a double-strand break in the specific region is executed by the Cas9. The recognition of the DNA sequence is done via a complementarity of a 20-nucleotide (nt) sequence within the sgRNA to the genomic target upstream of a 3-nt protospacer adjacent motif (PAM) at its 3' end. The CRISPR system is widely used in interrogating gene functions, and has wide applications in clinical detection, gene therapy, and agricultural improvement (6; 7; 8; 9; 10; 11).

Although the CRISPR/Cas9 system has many benefits, numerous studies have also shown that the mismatches between the designed sgRNA and DNA can be tolerated, resulting in the cleavage of unplanned genomic sites, termed off-target sites (OTS). Detecting OTS in CRISPR gene editing is highly important due to their potential disruptive effect, and designing sgRNAs with few OTS is desired (12; 13; 14). Many experimental techniques were developed to detect OTS, including GUIDE-seq, Digenome-Seq, HTGTS, SITE-Seq, CIRCLE-Seq, CHANGE-seq, Nuclea-seq, and BLISS (15; 16; 17; 18; 19; 20). However, these experimental techniques are limited in their ability to detect OTS with bulges, due to either experimental limits or inefficient post-processing pipelines.

Various computational methods have been developed to predict OTS based on experimental data. These methods can be split into three categories: alignment-based, ruled-based, and data-driven-based. Many data-driven methods are based on machine learning, including CRISTA, DeepCrispr, Elevation, among many others (21; 22; 23; 24; 25; 26; 27; 28). Several methods consider bulges, including CRISPR-Net and CRISPR-IP (26; 29). However, these methods are inaccurate in predicting OTS with bulges due to limited data of OTS with bulges.

Recently, new datasets of unprecedented scale and quality were produced for both OTS *in vitro* and *in cellula* (30). The CHANGE-seq dataset contains $202,040$ unique *in vitro* OTS across 110 sgRNAs, while the GUIDE-seq dataset contains $1,208$ unique *in cellula* OTS across 58 sgRNAs, which are a subset of the 110 sgRNAs in the CHANGE-seq dataset. But, while the CHANGE-seq dataset includes $23,692$ OTS with bulges, the GUIDE-seq datasets do not include any OTS with bulges. This critical bottleneck in the number of GUIDE-seq OTS with bulges prohibits the training of any machine-learning model to predict OTS *in cellula*.

To close this gap, in this study, we generated the most comprehensive dataset of OTS with bulges to date. We reprocessed the raw experimental GUIDE-seq data that is publicly available through the CHANGE-seq study. Through this processing, we both generated more than 450 novel OTS with bulges, and additional 20 new GUIDE-seq datasets. Using this dataset, we trained multiple machine-learning models to predict OTS with bulges. We evaluated these models on multiple datasets and demonstrated how they outperform the state-of-the-art by utilizing an embedding layer and transfer learning from *in vitro* to *in cellula* data. Based on the trained embedding layer, we produced novel visualizations of OTS representations with bulges, which recapitulate known and novel CRISPR/Cas9 biology.

Materials and Methods

A. Datasets and our updated preprocessing

CHANGE-seq and GUIDE-seq datasets. We used *in vitro* and *in cellula* high-throughput CRISPR OTS datasets (Table 1). CHANGE-seq is an advanced scalable, highly sensitive, and unbiased assay for measuring the genome-wide activity of CRISPR/Cas9 nucleases *in vitro* (30). The developers of CHANGE-seq applied their method to 110 sgRNA targets across human primary T cells. They preprocessed the experimental output to identify OTS with up to 1 DNA or RNA bulge and 4 mismatches compared to the sgRNA, or up to 6 mismatches and no bulges, which resulted in $202,040$ OTS. GUIDE-seq is an *in cellula* and unbiased approach for global detection of DNA double-stranded breaks introduced by CRISPR RNA-guided-nucleases (15). GUIDE-seq measurements are affected by chromatin accessibility and other epigenetic and cellular factors (31). The developers of CHANGE-seq conducted GUIDE-seq experiments to compare their measurements with the measurements produced by the CHANGE-seq assay.

Due to an outdated data post-processing pipeline, the published GUIDE-seq dataset lacked OTS with bulges. To address this gap, we applied an updated version of the GUIDE-seq pipeline (available at `https://github.com/tsailabSJ/guideseq`) to the raw data, which we downloaded from NCBI Sequence Read Archive (SRA) PRJNA625995, enabling us to identify a total of $1,520$ unique OTS for 58 sgRNAs, including 312 sites containing DNA or RNA bulges, where we set the same bulge and mismatches thresholds as were applied on the CHANGE-seq data. To remove duplicate OTS from the data, we averaged the read count of non-unique OTS over their replicates.

Our newly generated GUIDE-seq experiments. To apply the updated version of the GUIDE-seq pipeline, we had to acquire the raw high-throughput sequencing data files along with their corresponding DNA barcodes. While the sequencing data of the GUIDE-seq experiments is publicly available through NCBI SRA, the DNA barcodes were not published. Upon request, we obtained the sequencing barcode files from the researchers of the CHANGE-seq study who conducted and processed these GUIDE-seq experiments. However, we encountered inconsistencies between some of the barcode files and their raw sequencing data files.

To address these inconsistencies, we applied the GUIDE-seq pipeline using all available barcodes on each raw sequencing data file. While this process is computationally intensive, it confirms which barcodes are associated with each raw sequencing data file based on the demultiplexing step of the pipeline. We set the minimum number of reads for the demultiplexing step to the default value of $10,000$. Due to the unexpectedly large number of demultiplexed files (over 58 experiments and their replicates), we suspected that the raw data contained additional GUIDE-seq experiments. To investigate further, we examined if any of the 110 sgRNAs of the CHANGE-seq sgRNA library had one of the highest read counts in the sequencing data. In addition, due to the similarity among tested sgRNAs, we confirmed that each experiment was associated with at most one sgRNA.

During our analysis, we modified the GUIDE-seq pipeline to support parallel computing to speed up the process. The modified version is available at `https://github.com/ofiryaish/guideseq_parallelized`. Our GitHub repository, CRISPR-Bulge, contains the DNA barcodes in their correct association.

Generating active and inactive OTS for machine-learning models. An essential step in developing an effective off-target predictor is constructing datasets that include active and inactive OTS (i.e., sites obtained experimentally and potential sites based on sequence similarity that were not cleaved) (32). To obtain all potential OTS, we employed SWOffinder, a novel method we developed for searching CRISPR OTS with bulges (33). Compared to previous search methods such as CasOFFinder (34), which were utilized in CRISPR-Net and CRISPR-IP (26; 29), SWOffinder guarantees the detection of all potential OTS of Cas9 RNA-guided endonucleases, both with and without bulges, given a specific sgRNA and a reference genome. Based on the number of mismatches and bulges seen in the experimental data, we configured the search parameters to identify all potential OTS with up to 1 DNA or RNA bulge and 4 mismatches compared to the given sgRNA, or with up to 6 mismatches in the absence of bulges.

To generate the inactive OTS, we filtered out experimentally identified OTS from the pool of potential OTS obtained by SWOffinder. Specifically, since each off-target site can be uniquely identified by the quartet (sgRNA, chromosome, strand, and genome end position), we removed potential OTS from the list if their quartet identifier matched those of the experimentally identified OTS. Additionally, for each sgRNA, we excluded potential OTSs found within a genome window

Table 1. Datasets summary. Details of our processed datasets used for training and testing our machine-learning models.

Dataset	sgRNAs	Total active off-target sites	Active off-target sites with bulges*	Total inactive off-target sites	Inactive off-target sites with bulges*	Notes
CHANGE-seq	110	202,040	23,692	4,734,238	1,725,650	Active sites as obtained in the CHANGE-seq study (30)
GUIDE-seq	58	1,520	312	211,8820	739,258	1. Active sites with bulges obtained using our updated raw data reprocessing 2. Duplicate active sites were aggregated
Our newly generated GUIDE-seq experiments	20	646	147	1,150,063	438,228	New experiments obtained using our updated raw data reprocessing
TrueOT	32	230	10	1,495	197	Among experimentally validated sites, sites with an editing rate greater than 0.1% are considered active sites (35; 36)

*Off-target sites with up to 1 DNA or RNA bulge and 4 mismatches. Without bulge, sites with up to 6 mismatches are allowed

surrounding the experimentally identified sites for that particular sgRNA. This approach is not only based on the assumption that neighboring sites may be considered the same, especially when considering bulges, but also takes into account the fact that current state-of-the-art experimental techniques, such as CHANGE-seq and GUIDE-seq (15; 30), cannot precisely determine the exact position of an off-target site within a sequencing read. Similar to the GUIDE-seq pipeline, we defined the genome window as the identified off-target site plus an additional 25 flanking positions. We excluded those OTS from our dataset. Table 1 reports the total number of inactive OTS for the CHANGE-seq and GUIDE-seq datasets.

We defined experimentally obtained OTS as the active sets. In the CHANGE-seq study and our previous work (32), experimentally identified OTS cleaved *in vitro* with read counts below 100 were filtered out as noisy measurements. This filtering criterion resulted in a reduced size of the CHANGE-seq active set to $70,425$ OTS. Thus, we re-examined this filtering criterion.

The TrueOT benchmark. The TrueOT dataset (35; 36) is a comprehensive collection of experimental data designed specifically for evaluating potential off-target effects of CRISPR-induced DNA sequence alterations. The dataset comprises OTS obtained in 11 different studies. The inclusion criteria for the dataset required experiments to be performed in living cells to measure off-target editing efficiencies by directly assessing target sequence mutation rates and providing the chromosomal position of the OTS. The OTS were experimentally validated through next-generation sequencing of PCR amplicons, and OTS with an editing rate greater than 0.1% were labeled as active OTS. We excluded OTS that had both active and inactive labels, which are likely due to cell-type-specific factors. In addition, for each sgRNA, we removed duplicate OTS. As a result, our refined TrueOT benchmark dataset contains $1,725$ OTS, where 230 of them are active, and only 10 include bulges (Table 1).

B. Our deep neural networks to predict OTS

Sequence encoding. In our deep neural networks, the prediction of off-target activity is based on aligned sgRNA and off-target sequences, which include the PAM. To accommodate aligned sequences with bulges, we extend the encoding approach utilized in our previous study (32). Given an aligned sgRNA sequence S_t and an aligned off-target sequence S_o, we encode the pair of aligned nucleotides or gaps (represented by -) at position i in the aligned sequences into a 5×5 matrix $C_i(j_t, j_o)$ using the following scheme:

$$C_i(j_t, j_o) = \begin{cases} 1, & S_t[i] = V[j_t] \wedge S_o[i] = V[j_o],\ V = [A, C, G, T, -] \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where j_t and j_o range from 1 to 5, representing the four possible nucleotides and a gap. Prior to the encoding process, we adjust the N in the sgRNA PAM sequence to match the corresponding nucleotide in the off-target site.

To generate one-hot-encoding sequence features for a pair of aligned sequences, each matrix representing the aligned pair at a specific position is flattened. Subsequently, all of these vectors are stacked to form a one-hot encoding matrix. As the alignment length can be either 23 or 24, shorter alignments are left-padded with a pair of gaps. This results in a fixed-sized one-hot-encoding matrix of dimensions 25×24, which is then used as input to our models (Figure 1a).

Classification and regression models. We developed two types of predictive models for our study. The first, a classification model, aims to predict the cleavage probability of an off-target site. The second, a regression model, aims to predict the cleavage efficiency of an off-target site. In the case of a classification model, we defined the positive and negative sets as the off-target active and inactive sets. For the regression model labels, we employed a data transformation technique where the read counts, denoted as x, were log-transformed using

the formula $log(x+1)$ (32). We assigned inactive OTS a read count of 0 to reflect their lack of cleavage activity.

Deep neural network architectures. We trained three deep-neural-network architectures (inspired by the DeepHF study for the on-target prediction task (37)) for off-target classification and regression tasks, leveraging multilayer perceptron (MLP), Gated recurrent unit (GRU), and embedding techniques. The first architecture, which we denote by *MLP-Embd*, begins with an embedding layer of size 44, which transforms input sequences into dense representations. The embedded sequences are then processed by a flattening layer. Subsequently, the flattened features pass through two consecutive dense layers with 128 and 64 neurons with ReLU activation function. The last layer is a single neuron with a sigmoid or linear activation for the classification and regression tasks, respectively. The second architecture, which we denote *GRU-Embd*, closely resembles the MLP-Embd architecture. However, the embedded input sequences are fed into a GRU layer before being processed by the dense layers (Figure 1b). We set the number of GRU units to 64. The third architecture, which we denote *GRU*, is similar to GRU-Embd but without the embedding layer.

Model training and transfer learning to GUIDE-seq. To ensure consistency and prevent mixing datasets with different characteristics, we trained separate models for the CHANGE-seq and GUIDE-seq datasets. We trained the models using the Adam optimizer with a learning rate of 0.001. We used a batch size of 512 with a maximum number of training epochs set to 10. We utilized the binary cross-entropy and mean-squared-error loss functions for the classification and regression tasks, respectively. During each model training, we used a validation set comprising 10% of the training data to select the optimal epoch based on the validation loss. Considering the highly imbalanced nature of both the GUIDE-seq and CHANGE-seq datasets, we investigated the effectiveness of class weighting in neural network models, a technique we previously used in an XGBoost model (32). For each active or inactive set, we defined its weight to be proportional to the size of the other set. In addition to training a separate model solely on the GUIDE-seq data, we employed a transfer learning approach to utilize the large-scale CHANGE-seq dataset to train a dedicated model for the GUIDE-seq data. In this technique, we first trained our model on the CHANGE-seq data and then continued the training on the GUIDE-seq data.

C. Prediction performance evaluations

Evaluation metrics. We evaluated prediction performance over multiple sgRNAs. On the CHANGE-seq dataset, we randomly partitioned the dataset of 110 sgRNAs into 10 sets, each containing samples from 11 different sgRNAs. During each fold of a cross-validation process, one set was used as the test set for evaluation, while the remaining 9 sets were used as the training data. When evaluating the performance over the GUIDE-seq dataset, each train (test) fold included the GUIDE-seq experiments of the sgRNAs of the CHANGE-seq train (test) set experiments. We evaluated the performance of the models in the classification task using the area under the precision-recall curve (AUPR, calculated as average precision). This evaluation metric is particularly suitable for imbalanced datasets and is commonly used to evaluate OTS prediction performance. In addition, we defined a new evaluation by assessing performance on OTS with bulges. Unlike previous studies that primarily focused on the overall dataset performance, we sought to provide additional resolution to the evaluation of the predictive abilities of our models on the more challenging task of OTS with bulges.

Comparison to current methods for off-target prediction. For a baseline comparison, we implemented a modified version of the XGBoost model used in previous studies (30; 32) to enable prediction of OTS with bulges. The sequence encoding is our encoding where we flatten the one-hot-encoding matrix due to the limitation of XGBoost in processing two-dimensional data. To conduct more advanced comparisons (i.e., on the TrueOT dataset), we compared our models to the state-of-the-art methods that can handle OTS with bulges, CRISPR-Net (26) and CRISPR-IP (29).

D. Model interpretation

Visualizing the learned representation. To understand how the embedding models (i.e., MLP-Embd and GRU-Embd) capture the differences between the different types of OTS, we visualized the representation learned by the models. We randomly selected one sgRNA from the CHANGE-seq dataset and randomly generated $20,000$ OTS with up to 4 mismatches in the presence of one DNA/RNA bulge or up to 6 mismatches without bulges. We then employed T-distributed stochastic neighbor embedding (t-SNE) to project the high-dimensional embedded representations into a two-dimensional map (38). For comparison, we repeated this visualization analysis using the one-hot encoding sequence features.

Visualizing the positional effect on off-target activity. To visualize the positional effect of mismatches and bulges on off-target activity, we randomized 100 sgRNAs and, for each one, generated all possible OTS with a single mismatch or a DNA/RNA bulge, and all double-mismatched OTS. To assess the positional effect, we averaged the predictions across the 100 sgRNAs for each type of edit (mismatch, DNA/RNA bulge, and double mismatch) and plotted the average prediction as a function of the edit position. We performed this analysis on both the CHANGE-seq and GUIDE-seq models to assess the consistency between the principles learned from *in vitro* and *in cellula* data.

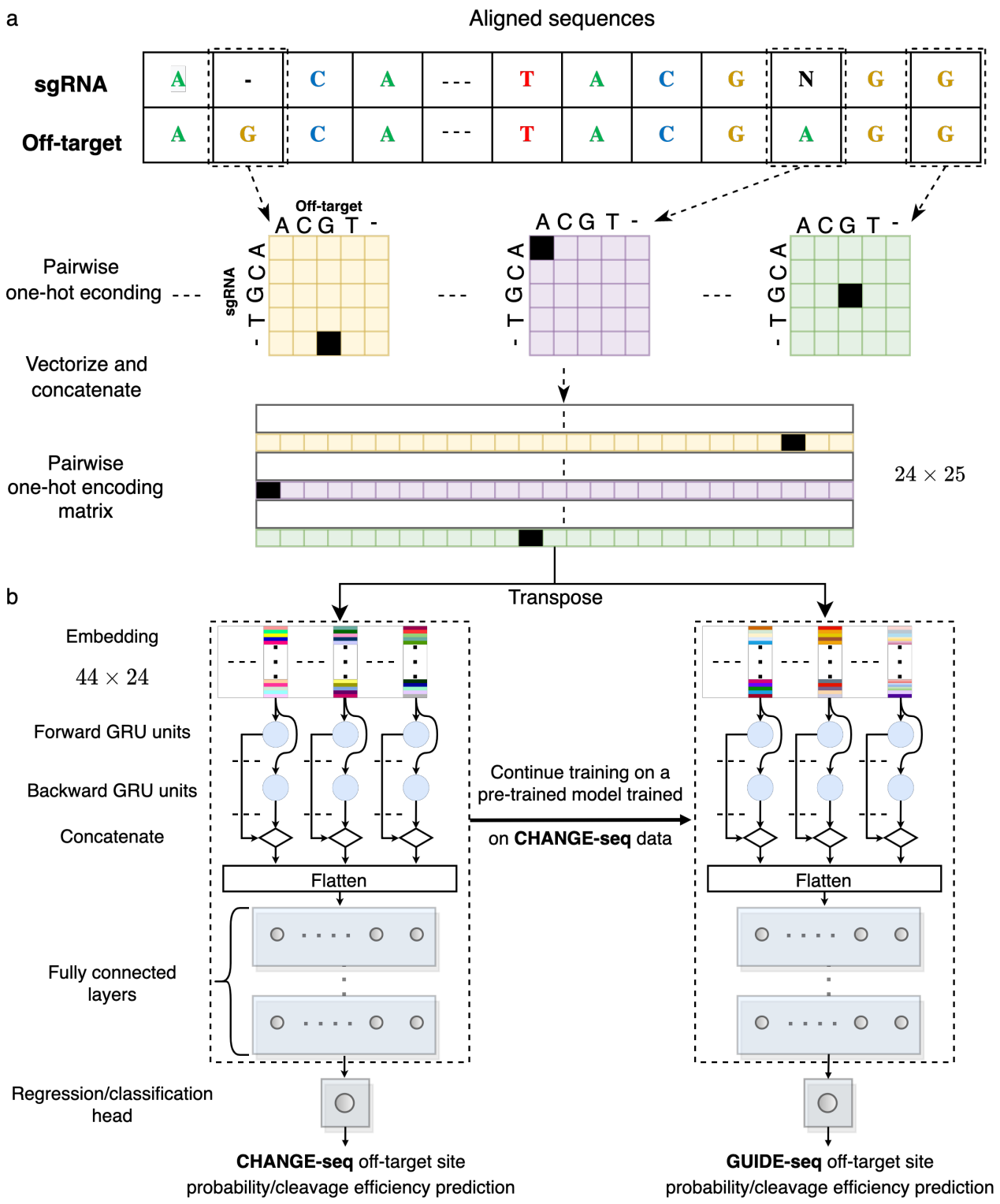

Figure 1. Predictive model overview. (a-b) Illustration of the GRU-Embd architecture and its training process. Given an aligned pair of sgRNA and off-target site sequences, the sequences are processed using pairwise one-hot encoding, and then the flattened positional vectors are concatenated (a). Subsequently, the fixed-sized one-hot-encoding matrix is provided as input to the GRU-Embd classifier/regressor to predict the off-target site probability/cleavage efficiency (b). A transfer learning approach utilizes the large-scale CHANGE-seq dataset to train a dedicated model for the GUIDE-seq data.

Results

E. Our new GUIDE-seq datasets including hundreds of CRISPR off-target sites with bulges

We generated the largest dataset to date of CRISPR OTS with bulges (Table 1). We retrieved data for all 58 originally published GUIDE-seq experiments, extracted novel OTS with bulges, and generated 20 new GUIDE-seq experiments. As a result, we expanded the set of GUIDE-seq datasets to cover 78 sgRNAs, with a

combined total of $2,166$ identified OTS, including 459 OTS with bulges. To the best of our knowledge, to this date, this is the most comprehensive set of GUIDE-seq datasets containing OTS with bulges.

F. Improved prediction of CRISPR off-target sites with bulges

We first explored the effect of removing the read-count threshold and using class weighting in the training process of our neural network models to train on all available data. We evaluated the models using a leave-11-sgRNAs-out cross-validation. To perform an objective comparison when examining the removal of the read-count threshold and for consistency with previous studies, we evaluated the performance on the subset of OTS with a read-count threshold of 100.

The models exhibit improved performance when the read-count threshold is removed (Figures 2a-b). For instance, in the regression models, the GRU model without class weighting and thresholding achieves an average AUPR of 0.64, whereas the comparable model with thresholding attains an average AUPR of 0.601 (p-value $= 9.76 \times 10^{-3}$ via Wilcoxon signed-rank test). Thus, we proceeded with removing the read-count threshold in all model training procedures. The full results, on each test fold, are reported in Supplementary Table S1.

We observed different effects of class weighting for the XGBoost and neural network models. Consistent with our previous study (32), the XGBoost model benefits from class weighting in all model variations. However, when examining the neural network models, we find no such improvement and even a decrease in the average AUPR for the regression models. For example, the GRU-Embd model without class weighting and without thresholding achieves an average AUPR of 0.628, while the comparable model with class weighting obtains an average AUPR of 0.604 (p-value $= 0.049$). Thus, for our subsequent regression model evaluations, we exclude class weighting from the training process. Due to the modest improvement in the average AUPR, we incorporated class weighting into the model training of the classification models. The full results, on each test fold, are reported in Supplementary Table S1.

G. Unique evaluation of CRISPR off-target sites with bulges

We uniquely compared the prediction performance of the classification and regression models on a subset of OTS consisting exclusively of OTS with bulges as well as on the entire dataset. We focused specifically on the GRU and GRU-Embd models, as the results indicated their superiority over the XGBoost and MLP-Embd models (Figures 2a-b). We computed the statistical significance based only on 10 AUPRs.

When evaluating the performance on the entire CHANGE-seq dataset (Figure 2c), the regression models consistently achieved higher average AUPR values compared to their corresponding classification models (albeit without statistical significance). For instance, the best regression model, GRU, obtained an average AUPR of 0.64, while the top classification model, GRU with class weighting, achieved an average AUPR of 0.624 (p-value $= 0.492$).

When evaluating the performance on the subset containing OTS with bulges (Figure 2d), the improvement of the best regression model was more prominent and statistically significant. The GRU regression model achieved the highest average AUPR of 0.476, whereas its corresponding classification model with class weighting obtained an average AUPR of 0.432 (p-value $= 1.37 \times 10^{-2}$). Based on these findings, we continued with the regression models for further analysis. The full results, on each test fold, are reported in Supplementary Table S2.

H. Improved prediction of CRISPR off-target sites with bulges *in cellula* through transfer learning

Following the training on the CHANGE-seq dataset (i.e., *in vitro* data), we fine-tuned the regression models to *in cellula* data by continuing the training on the GUIDE-seq data. To assess the effectiveness of our transfer learning, we compared its performance with models trained solely on GUIDE-seq or CHANGE-seq data. The GUIDE-seq data we used for training and testing the models in this evaluation is the subset of 58 sgRNAs designed originally in the CHANGE-seq study.

In both evaluations on the entire and only-bulges GUIDE-seq datasets, we observed that the transfer learning approach outperforms the other models (Figures 3a-b). For instance, on the entire GUIDE-seq dataset evaluation (Figure 3a), the best performing model is the GRU trained with transfer learning, which achieved an average AUPR of 0.743 compared to 0.586 and 0.260 obtained by its corresponding GRU models trained solely on GUIDE-seq and CAHNGE-seq data, respectively (p-values $< 9.77 \times 10^{-3}$). In the evaluation on the only-bulges dataset (Figure 3b), the GRU model also showed the best performance with an average AUPR of 0.438, compared to its corresponding GRU models trained solely on GUIDE-seq and CHANGE-seq data, which resulted in average AUPRs of 0.211 and 0.151, respectively (p-values $< 1.57 \times 10^{-2}$). The full results, on each test fold, are reported in Supplementary Table S3.

In our evaluations on the only-bugles GUIDE-seq dataset, the average AUPR was computed using only 7 AUPR scores instead of 10 since some of the cross-validation folds did not contain any active OTS with bulges. As a result, statistical significance was not achieved for all comparisons despite the large differences. Similar to the performance on the CHANGE-seq dataset, the GRU model achieved the best average AUPR among our transfer-learning models. The full results, on each test fold, are reported in Supplementary Table S3.

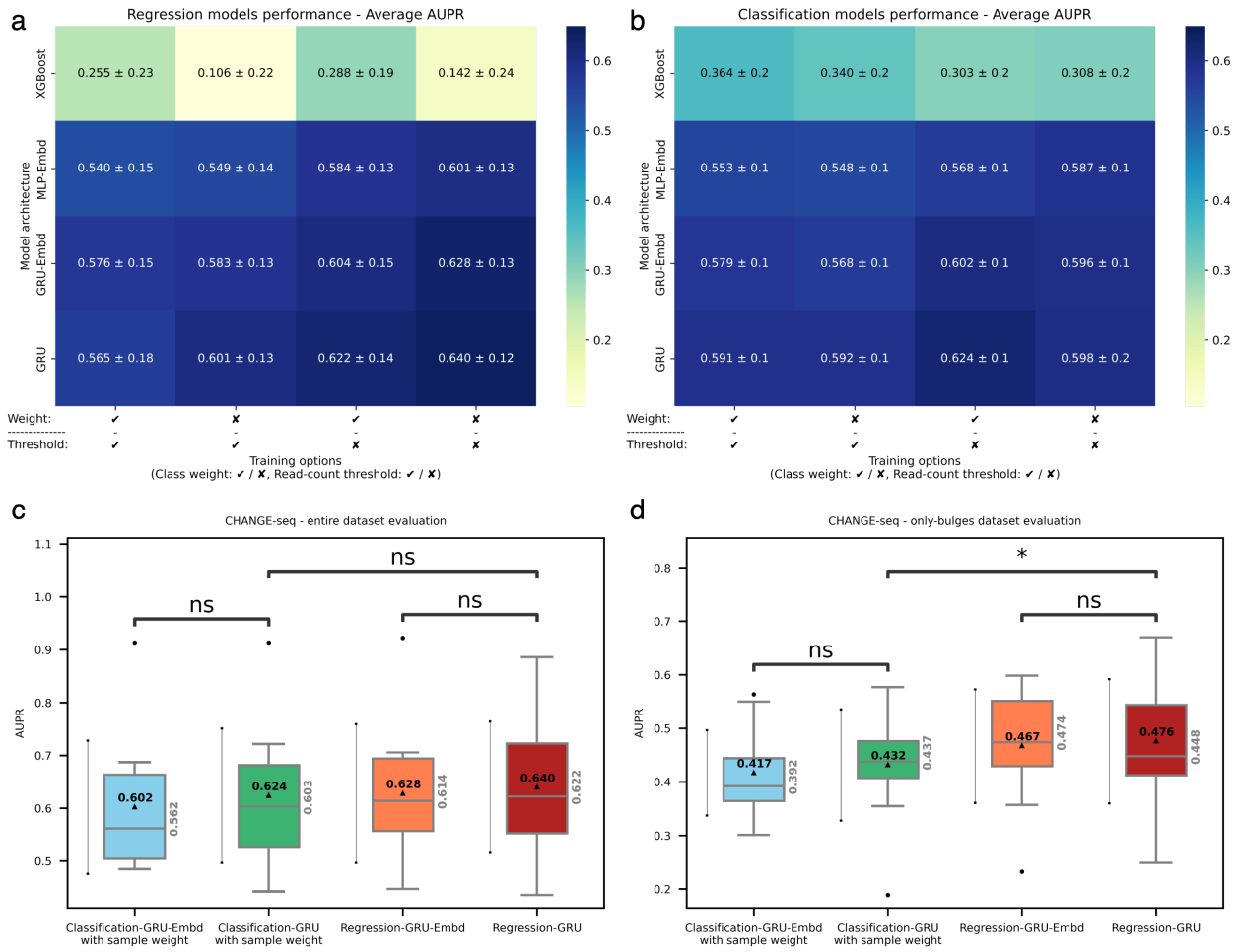

Figure 2. Prediction performance evaluations on *in vitro* data. (a-b) A comparison of the XGBoost, MLP-Embd, GRU, GRU-Embd regression (a) and classification (b) models on the CHANGE-seq data examining the effect of using class weighting and read-count threshold. We gauged the prediction performance of the models by average AUPR over the leave-11-sgRNAs-out cross-validation folds. **(a-b)** A comparison of the GRU and GRU-Embd regression and classification models examining performance on the CHANGE-seq dataset (c) and its only-bulges subset (d). The average AUPR values are denoted by a triangle within each box plot. An Error bar of one standard deviation is located on the left of each box plot. The median AUPR values are reported in grey on the right of each box plot. Statistical significance via Wilcoxon signed-rank test denoted by: ns $5 \cdot 10^{-2} < p \leq 1$; * $1 \cdot 10^{-2} < p \leq 5 \cdot 10^{-2}$.

I. Evaluation of our newly generated GUIDE-seq datasets

To validate the quality of our 20 newly generated GUIDE-seq experiments, we tested the performance of our transfer learning models on them. We trained the pre-trained models on the entire CAHGNE-seq dataset, excluding the sgRNAs of the new 20 GUIDE-seq datasets. Then, we continued the training on the subset of 58 GUIDE-seq sgRNAs that were originally published in the CHANGE-seq study. The prediction performance on the 20 new GUIDE-seq experiments is on par with the performance obtained on the 58 original GUIDE-seq experiments (Figures 3c-d), indicating the high quality of the new dataset. On the entire new GUIDE-seq dataset (Figure 3c), the GRU and GRU-Embd achieved an AUPR of 0.671 and 0.668, respectively, outperforming the MLP-Embd model, which achieved an AUPR of 0.618. Similarly, the GRU-Embd and GRU resulted in better prediction performance on the only-bulges evaluation, achieving an AUPR of 0.487 and 0.471, where the MLP-Embd achieved only an AUPR of 0.399 (Figure 3d).

J. Benchmarking methods on an independent experimentally verified off-target site dataset

In our final prediction performance evaluation, we evaluated the prediction performance of our models trained on the entire CHANGE-seq and GUIDE-seq datasets through transfer learning on experimentally validated true OTS generated using next-generation sequencing on PCR amplicons techniques, such as rhAmpSeq (39), which are considered the gold standard for validating true OTS (35). Our refined TrueOT benchmark dataset contains $1,725$ OTS, 230 of which are active, including 10 with bulges.

Our GRU-based and MLP-based models trained with our novel dataset exhibited superior performance when compared to state-of-the-art prediction methods,

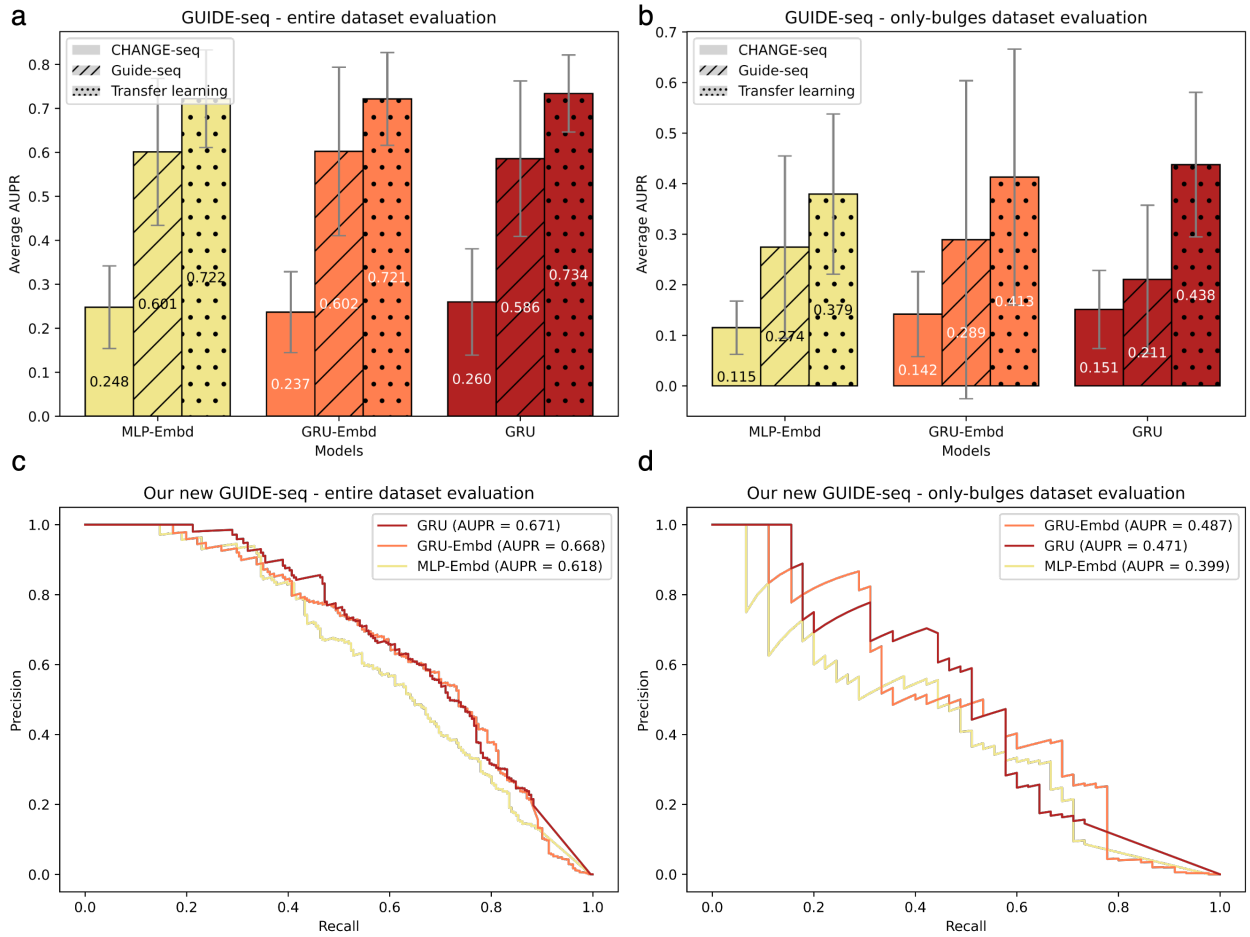

Figure 3. Prediction performance evaluations on *in cellula* data. (a-b) Evaluation on the entire GUIDE-seq dataset published in the CHANGE-seq study (a) and its only-bulges subset (b). MLP-Embd, GRU, and GRU-Embd regression models trained on different data are examined: trained only on CHANGE-seq, trained only on GUIDE-seq, and trained on CHANGE-seq and then fine-tuned on GUIDE-seq (denoted as Transfer learning). Due to the limited number of OTS with bulges, in the only-bulges subset evaluation, the Average AUPR was computed over 7 test folds instead of 10 test folds used in the entire dataset evaluation. **(c-d)** Evaluation of the transfer-learning-based MLP-Embd, GRU, and GRU-Embd regression models on our newly generated 20 GUIDE-seq datasets (c) and their only-bulges subset (d).

CRISPR-Net and CRISPR-IP (26; 29), with the most superior advantage obtained in the evaluation of the subset containing OTS with only bulges (Figures 4a-b). On the entire TrueOT benchmark dataset (Figure 4a), the GRU model achieved the best AUPR of 0.635, outperforming GRU-Embd, MLP-Embd, CRISPR-Net, and CRISPR-IP models which achieved an AUPR of 0.612, 0.584, 0.535, and 0.482, respectively. Similarly, on the only-bulges evaluation, the GRU obtained the best AUPR of 0.491 compared to the models MLP-Embd, GRU-Embd, CRISPR-Net, and CRISPR-IP which achieved an AUPR of 0.463, 0.450, 0.270, and 0.141, respectively. The same trends are observed when evaluating using the AUROC metric (Supplementary Figure S1). The same trends are also observed when removing from our refined TrueOT dataset experiments on which CRISPR-Net was trained (Supplementary Figure S2).

K. Visualization of the learned embedded representation of off-target sites with bulges

To understand what the network has learned, we visualized the output of the embedding layer for various OTS in a t-SNE plot. Figure 5 visualizes the learned embedded representation by the GRU-Embd model projected into a two-dimensional map using t-SNE. The visualization effectively showcases a clear separation of OTS based on the presence or absence of bulges. Moreover, it distinctly illustrates the separation between OTS, featuring different types of bulges.

The observed separation in the learned embedded space can give us insights into the model prediction preferences assuming that proximity in the learned embedded space implies similar predictions by the model. While the model treats OTS without bulges similarly as their cluster is very dense, it treats OTS with bulges differently according to bulge type and location. Generally, OTS with bulges positioned farther from the PAM exhibit closer proximity to the cluster of OTS without bulges. We conclude that OTS with bulges close to the PAM

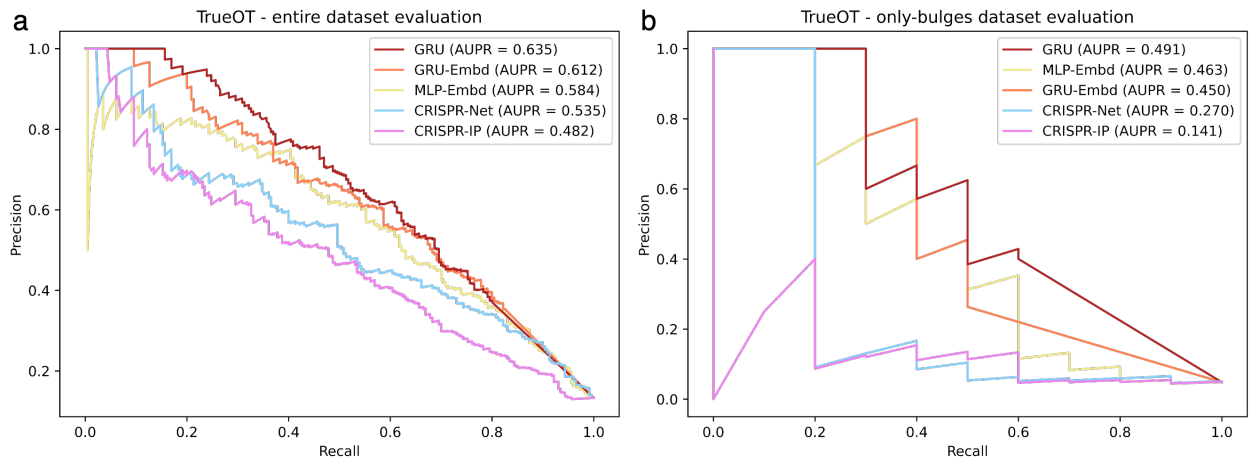

Figure 4. Prediction performance evaluations on an experimentally verified dataset. (a-b) Comparison of our regression models MLP-Embd, GRU, and GRU-Embd trained on the entire CHANGE-seq and GUIDE-seq data through transfer learning with state-of-the-art-models CRISPR-Net and CRISPR-IP on our refined TrueOT benchmark dataset (a) and its only-bulges subset (b).

are less likely to be active, reaffirming previous findings obtained by synthetic assays (40; 41). The importance of the learned representation by an embedding layer is emphasized by comparison to a similar plot visualized based on one-hot-encoding sequence features, where these trends are not as apparent (Supplementary Figure S3).

L. Analyzing positional effect on off-target activity

We analyzed the positional effect of mismatches and bulges on off-target activity learned by our GRU model in the case of *in vitro* and *in cellula* data. We examined the positional effect of a single edit compared to the sgRNA (Figure 6a-f). In general, as expected, the models for *in vitro* and *in cellula* data learn that edits near the PAM are unlikely in active OTS. As an exception, the GRU *in vitro* model learns a pattern where a single mismatch and RNA bulge around locations 7 and 15 are less tolerant to edits compared to adjacent positions (Figures 6a,c). Remarkably, these position-dependent patterns align with findings from previous synthetic assays (40). The *in cellula* GRU model also captured similar patterns, albeit with reduced intensity (Figures 6b,d). Another dissimilarity between the *in vitro* and *in cellula* models is the higher sensitivity of the *in cellula* model to RNA bulges near the PAM (Figures 6c-d).

Furthermore, we analyzed the positional effect of two mismatches compared to the sgRNA (Figure 6g-h). Overall, we see that the general patterns learned in the single mismatch case align with the trends here. However, in contrast to the single mismatch scenario, where we observed a similar trend learned by the GRU models for both *in vitro* or *in cellula*, we noticed that the *in cellula* model exhibited increased intolerance to more than one mismatch, even in positions distant from the PAM. Similar trends were observed when performing the same analysis on the GRU-Embd model (Supplementary Figure S3).

Discussion

In this study, we made several contributions to the CRISPR/Cas9 off-target research field. First, to the best of our knowledge, we generated the largest dataset to date of OTS with bulges. We generated new in cellula data, which includes hundreds of sites with bulges, an order of magnitude more than previous datasets. Second, using our newly generated dataset, we trained deep neural networks to predict off-target activity by utilizing embedding and recurrent neural networks with transfer-learning techniques to benefit from both *in vitro* and in cellula data. Our networks outperformed the state of the art by a large margin on experimentally validated true OTS data, which is considered the gold standard for validating true OTS (35). Third, we defined a new metric to evaluate models for off-target prediction with bulges, focusing on active and inactive OTS with bulges. This allows the close inspection of methods' performance when considering bulges, which may be obscured when evaluating on all the data. Fourth, thanks to the embedding layer, we developed a new visualization technique that utilizes the learned representation of OTS with bulges and identified their relevance and importance as learned by the model. Fifth, we analyzed the positional effect of mismatches and bulges on the off-target activity learned by our model, reaffirming previous knowledge and providing new insights.

Our study has several limitations. First, while we made a breakthrough in the scale of GUIDE-seq datasets, and in the scale of OTS with bulges in them, this data is still limited compared to other genomic tasks, where tens of thousands of measurements are available. We tackled this limit by transfer learning from the CHANGE-seq dataset, but more GUIDE-seq data is expected to improve the prediction performance of our models. Second, while we evaluated and compared our models to the state of the art on the most comprehensive dataset of experimentally verified OTS, this data still needs to

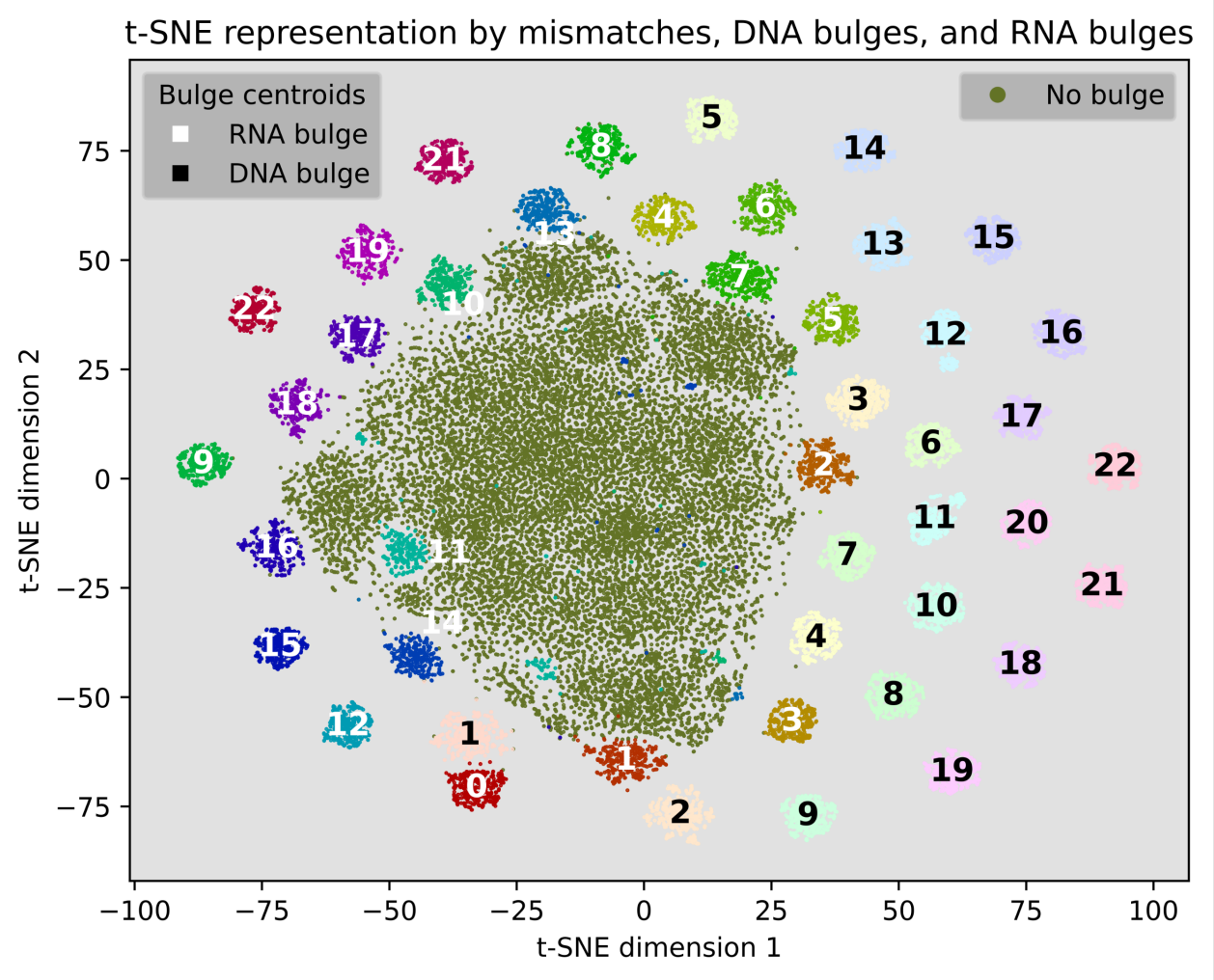

Figure 5. Visualization of the learned embedded representation of off-target sites with bulges. Learned embedded representation by the GRU-Embd model projected into a two-dimensional map using t-SNE. Comparing OTS with mismatches to OTS with bulges according to their bulge type and location. The centroid of OTS with the same bulge location is illustrated in each map by the position of the label.

be improved in scale, and additional low-throughput experimental validation of specific OTS is needed for robust benchmarking. Third, a prominent limitation lies in the dependence of our method and data-generation pipelines on the optimal alignment between the sgRNA and the off-target site. This alignment has a critical effect on both the data generation as the pipeline identifies optimal sites in a window, and the machine-learning model as it receives the alignment as input.

Several aspects of our study require further research in the future. First, in addition to the data we used, various experimental assays developed to measure OTS in high-throughput were not used in this study (40; 42; 43). Integrating them, either by data combination or by a meta-predictor that considers the predictions of models trained on these data may improve prediction performance. Second, the efficiency of assays for measuring the genome-wide activity of CRISPR/Cas9 nucleases *in cellula* is limited by chromatin accessibility and other epigenetic factors (31; 44; 45). Therefore, including such information in our models may provide more insights regarding their potential for future improvements. Third, current methods to predict OTS assume an alignment between the sgRNA and the off-target, but multiple alignments may exist. Having a method that receives as input multiple possible alignments, or no alignment at all, may improve prediction performance and shed more light on the interplay between sgRNAs and their OTS.

Conclusion

We presented a groundbreaking contribution to the CRISPR/Cas9 gene-editing field by introducing the most comprehensive dataset of OTS with bulges. By meticulously curating and analyzing a vast collection of experimental data, the study sheds new light on the intricate mechanisms underlying off-target effects induced by CRISPR/Cas9. The inclusion of bulges expands our understanding of the complex interplay between guide RNA and target DNA, providing insights into the factors influencing off-target activity. The availability of this extensive dataset will significantly benefit the scientific community, enabling researchers to develop more accurate prediction models and refine the

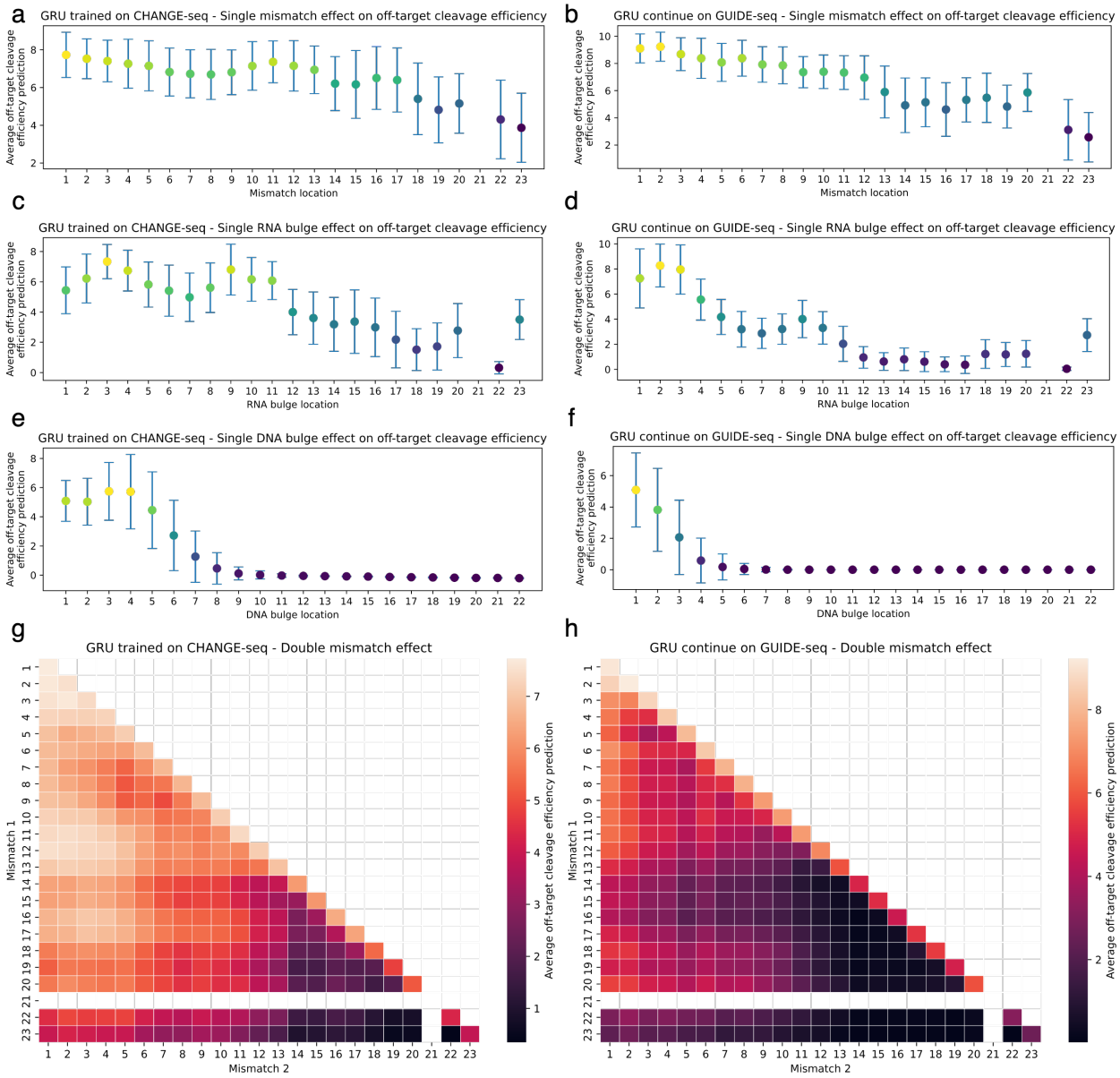

Figure 6. Positional effect of mismatches and bulges on off-target activity learned by our GRU model. a-f Positional effect of a single mismatch (a-b), RNA bulge (c-d), or DNA bulge (e-f) compared to the sgRNA predicted by a model for *in vitro* (a, c, and e) and *in cellula* (b, d, and f) data. **g-h** Positional effect of two mismatches compared to the sgRNA predicted by a model for *in vitro* (g) and *in cellula* (h) data.

design of CRISPR/Cas9-based therapies. Ultimately, this research empowers the ongoing efforts to enhance the precision, safety, and efficacy of CRISPR/Cas9 technology.

Competing interests

The authors declare they have no competing interests.

Author contributions statement

OY conducted all the analyses and developed all the codes used in this study. YO supervised this study. OY and YO wrote the manuscript.

Acknowledgements

We thank our colleagues at the CRISPR-IL consortium for thoughtful discussions over the off-target prediction problem. Ofir acknowledges the support of the Azrieli Ph.D. fellowship, Nehemia Levtzion Ph.D. fellowship from the Israeli Council for Higher Education, Kretiman Ph.D. fellowship at Ben-Gurion University, and the cloud-compute credit by the Planning and Budgeting Committee through IUCC.

Funding

The Israel Innovation Authority supported the study through the CRISPR-IL consortium. This research was partially supported by the Israel Science Foundation

(grant no. 358/21).

Data Availability

The software and code are publicly available via github.com/OrensteinLab/CRISPR-Bulge.

Supplemental Data

M. Supplementary Figures

Supplementary Figures S1-4.

N. Supplementary Tables

Supplementary Tables S1-3.

Bibliography

1. L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, et al., Multiplex genome engineering using CRISPR/Cas systems, Science 339 (6121) (2013) 819–823.
2. F. A. Ran, P. D. Hsu, C.-Y. Lin, J. S. Gootenberg, S. Konermann, A. E. Trevino, D. A. Scott, A. Inoue, S. Matoba, Y. Zhang, et al., Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity, Cell 154 (6) (2013) 1380–1389.
3. J. A. Doudna, E. Charpentier, The new frontier of genome engineering with CRISPR-Cas9, Science 346 (6213) (2014).
4. R. Barrangou, J. A. Doudna, Applications of CRISPR technologies in research and beyond, Nature Biotechnology 34 (9) (2016) 933–941.
5. M. Adli, The CRISPR tool kit for genome editing and beyond, Nature Communications 9 (1) (2018) 1–13.
6. H. Yang, H. Wang, C. S. Shivalila, A. W. Cheng, L. Shi, R. Jaenisch, One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering, Cell 154 (6) (2013) 1370–1379.
7. F. J. Najm, C. Strand, K. F. Donovan, M. Hegde, K. R. Sanson, E. W. Vaimberg, M. E. Sullender, E. Hartenian, Z. Kalani, N. Fusi, et al., Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens, Nature Biotechnology 36 (2) (2018) 179.
8. T. Gonatopoulos-Pournatzis, M. Aregger, K. R. Brown, S. Farhangmehr, U. Braunschweig, H. N. Ward, K. C. Ha, A. Weiss, M. Billmann, T. Durbic, et al., Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9–Cas12a platform, Nature Biotechnology 38 (5) (2020) 638–648.
9. C. Myhrvold, C. A. Freije, J. S. Gootenberg, O. O. Abudayyeh, H. C. Metsky, A. F. Durbin, M. J. Kellner, A. L. Tan, L. M. Paul, L. A. Parham, et al., Field-deployable viral diagnostics using CRISPR-Cas13, Science 360 (6387) (2018) 444–448.
10. G. J. Knott, J. A. Doudna, CRISPR-Cas guides the future of genetic engineering, Science 361 (6405) (2018) 866–869.
11. J. Shi, H. Gao, H. Wang, H. R. Lafitte, R. L. Archibald, M. Yang, S. M. Hakimi, H. Mo, J. E. Habben, ARGOS 8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions, Plant Biotechnology Journal 15 (2) (2017) 207–216.
12. V. Pattanayak, S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna, D. R. Liu, High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity, Nature Biotechnology 31 (9) (2013) 839–843.
13. Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung, J. D. Sander, High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells, Nature Biotechnology 31 (9) (2013) 822–826.
14. S. W. Cho, S. Kim, Y. Kim, J. Kweon, H. S. Kim, S. Bae, J.-S. Kim, Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases, Genome Research 24 (1) (2014) 132–141.
15. S. Q. Tsai, Z. Zheng, N. T. Nguyen, M. Liebers, V. V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A. J. Iafrate, L. P. Le, et al., GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases, Nature Biotechnology 33 (2) (2015) 187–197.
16. D. Kim, S. Bae, J. Park, E. Kim, S. Kim, H. R. Yu, J. Hwang, J.-I. Kim, J.-S. Kim, Digenome-seq: genome-wide profiling of crispr-cas9 off-target effects in human cells, Nature Methods 12 (3) (2015) 237–243.
17. R. L. Frock, J. Hu, R. M. Meyers, Y.-J. Ho, E. Kii, F. W. Alt, Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases, Nature Biotechnology 33 (2) (2015) 179–186.
18. P. Cameron, C. K. Fuller, P. D. Donohoue, B. N. Jones, M. S. Thompson, M. M. Carter, S. Gradia, B. Vidal, E. Garner, E. M. Slorach, et al., Mapping the genomic landscape of CRISPR–Cas9 cleavage, Nature Methods 14 (6) (2017) 600–606.
19. S. Q. Tsai, N. T. Nguyen, J. Malagon-Lopez, V. V. Topkar, M. J. Aryee, J. K. Joung, CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets, Nature Methods 14 (6) (2017) 607.
20. W. X. Yan, R. Mirzazadeh, S. Garnerone, D. Scott, M. W. Schneider, T. Kallas, J. Custodio, E. Wernersson, Y. Li, L. Gao, et al., BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks, Nature Communications 8 (1) (2017) 1–9.
21. S. Abadi, W. X. Yan, D. Amar, I. Mayrose, A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action, PLoS Computational Biology 13 (10) (2017) e1005807.
22. Q. Liu, X. Cheng, G. Liu, B. Li, X. Liu, Deep learning improves the ability of sgRNA off-target propensity prediction, BMC Bioinformatics 21 (1) (2020) 1–15.
23. G. Chuai, H. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan, et al., DeepCRISPR: optimized CRISPR guide RNA design by deep learning, Genome Biology 19 (1) (2018) 1–18.
24. J. Listgarten, M. Weinstein, B. P. Kleinstiver, A. A. Sousa, J. K. Joung, J. Crawford, K. Gao, L. Hoang, M. Elibol, J. G. Doench, et al., Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs, Nature Biomedical Engineering 2 (1) (2018) 38–47.
25. J. Lin, K.-C. Wong, Off-target predictions in CRISPR-Cas9 gene editing using deep learning, Bioinformatics 34 (17) (2018) i656–i663.
26. J. Lin, Z. Zhang, S. Zhang, J. Chen, K.-C. Wong, CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels, Advanced Science 7 (13) (2020) 1903562.
27. Y. Zhang, Y. Long, R. Yin, C. K. Kwoh, DL-CRISPR: A Deep Learning Method for Off-Target Activity Prediction in CRISPR/Cas9 With Data Augmentation, IEEE Access 8 (2020) 76610–76617.
28. J. Charlier, R. Nadon, V. Makarenkov, Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-Cas9 gene editing, Bioinformatics (Oxford, England) (2021) btab112.
29. Z.-R. Zhang, Z.-R. Jiang, Effective use of sequence information to predict crispr-cas9 off-target, Computational and Structural Biotechnology Journal 20 (2022) 650–661.
30. C. R. Lazzarotto, N. L. Malinin, Y. Li, R. Zhang, Y. Yang, G. Lee, E. Cowley, Y. He, X. Lan, K. Jividen, et al., CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity, Nature Biotechnology 38 (11) (2020) 1317–1327.
31. P. Akcakaya, M. L. Bobbin, J. A. Guo, J. Malagon-Lopez, K. Clement, S. P. Garcia, M. D. Fellows, M. J. Porritt, M. A. Firth, A. Carreras, et al., In vivo CRISPR editing with no detectable genome-wide off-target mutations, Nature 561 (7723) (2018) 416–419.
32. O. Yaish, M. Asif, Y. Orenstein, A systematic evaluation of data processing and problem formulation of crispr off-target site prediction, Briefings in Bioinformatics 23 (5) (2022) bbac157.
33. O. Yaish, A. Malle, E. Cohen, Y. Orenstein, Swoffinder, https://github.com/OrensteinLab/SWOffinder (2023).
34. S. Bae, J. Park, J.-S. Kim, Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases, Bioinformatics 30 (10) (2014) 1473–1475.
35. X. R. Bao, Y. Pan, C. M. Lee, T. H. Davis, G. Bao, Tools for experimental and computational analyses of off-target editing by programmable nucleases, Nature Protocols 16 (1) (2021) 10–26.
36. P. K. Kota, Y. Pan, H.-A. Vu, M. Cao, R. G. Baraniuk, G. Bao, The need for transfer learning in crispr-cas off-target scoring, bioRxiv (2021) 2021–08.
37. D. Wang, C. Zhang, B. Wang, B. Li, Q. Wang, D. Liu, H. Wang, Y. Zhou, L. Shi, F. Lan, et al., Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning, Nature Communications 10 (1) (2019) 1–14.
38. L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of Machine Learning Research 9 (11) (2008).
39. M. Schubert, G. R. Rettig, G. Kurgan, M. S. McNeill, J. Wang, S. Fiedler, J. DeWitt, M. A. Collingwood, C. Vakulskas, M. A. Behlke, Evaluate CRISPR-Cas9 Edits Quickly and Accurately with rhAmpSeq Targeted Sequencing.
40. S. K. Jones, J. A. Hawkins, N. V. Johnson, C. Jung, K. Hu, J. R. Rybarski, J. S. Chen, J. A. Doudna, W. H. Press, I. J. Finkelstein, Massively parallel kinetic profiling of natural and engineered crispr nucleases, Biophysical Journal 120 (3) (2021) 138a.
41. E. A. Boyle, W. R. Becker, H. B. Bai, J. S. Chen, J. A. Doudna, W. J. Greenleaf, Quantification of cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement, Science Advances 7 (8) (2021) eabe5496.
42. R. Fu, W. He, J. Dou, O. D. Villarreal, E. Bedford, H. Wang, C. Hou, L. Zhang, Y. Wang, D. Ma, et al., Systematic decomposition of sequence determinants governing crispr/cas9 specificity, Nature communications 13 (1) (2022) 474.
43. H. Zhang, J. Yan, Z. Lu, Y. Zhou, Q. Zhang, T. Cui, Y. Li, H. Chen, L. Ma, Deep sampling of grna in the human genome and deep-learning-informed prediction of grna activities, Cell Discovery 9 (1) (2023) 48.
44. J. K. Mak, F. Störtz, P. Minary, Comprehensive computational analysis of epigenetic descriptors affecting crispr-cas9 off-target activity, BMC Genomics 23 (1) (2022) 805.
45. F. Störtz, J. K. Mak, P. Minary, picrispr: Physically informed deep learning models for crispr/cas9 off-target cleavage prediction, Artificial Intelligence in the Life Sciences 3 (2023) 100075.

Supplementary Figures

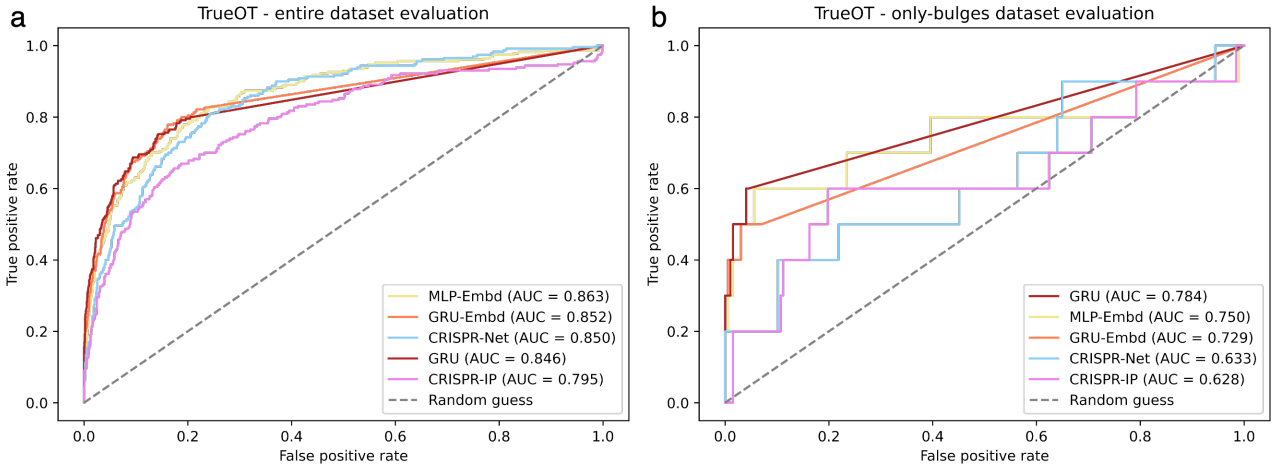

Figure S1. Prediction performance evaluations on an experimentally verified dataset. (a-b) Comparison of our regression models MLP-Embd, GRU, and GRU-Embd trained on the entire CHANGE-seq and GUIDE-seq data through transfer learning with state-of-the-art-models CRISPR-Net and CRISPR-IP on our refined TrueOT benchmark dataset (a) and its only-bulges subset (b).

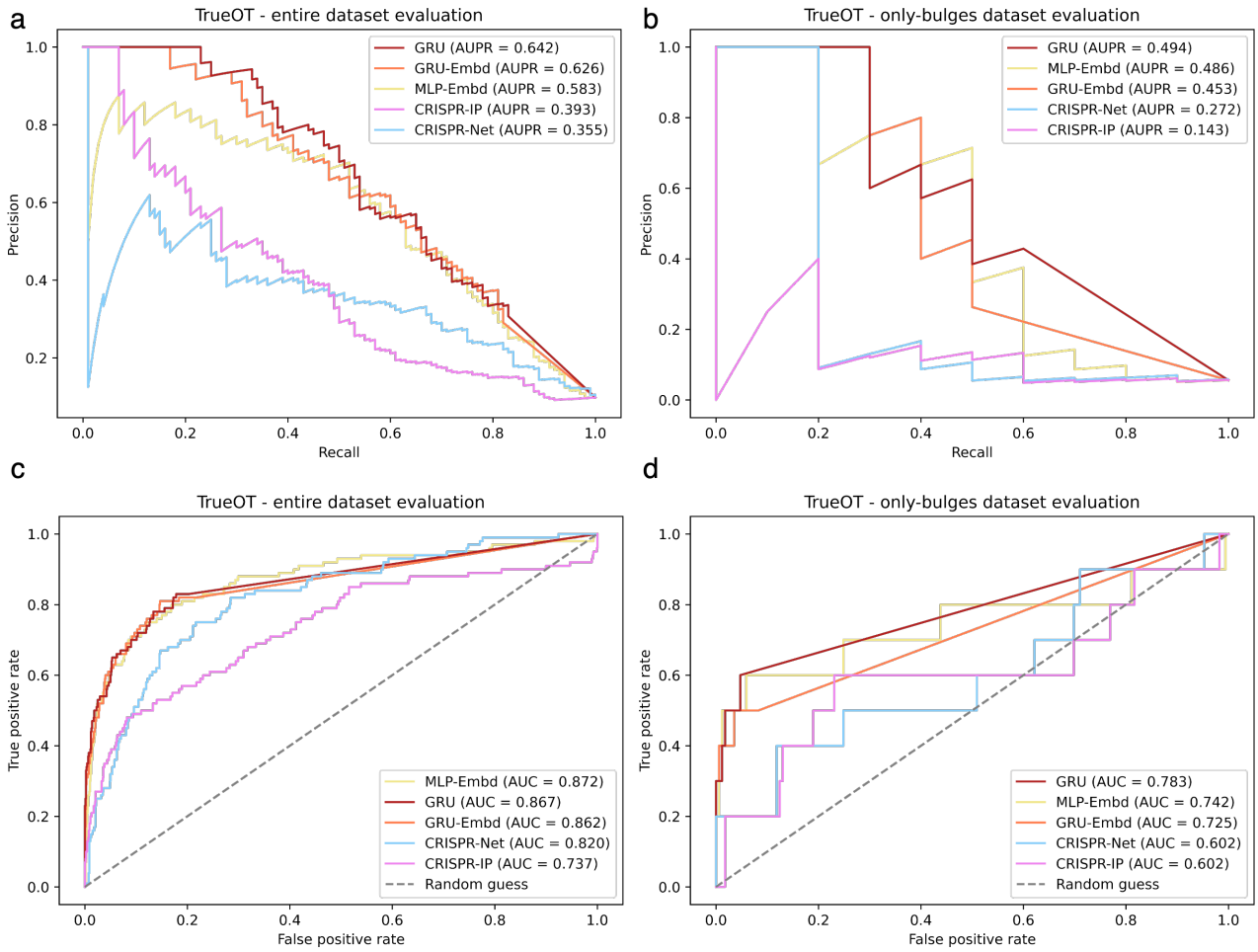

Figure S2. Prediction performance evaluations on an experimentally verified dataset after removing experiments on which CRISPR-Net was trained. (a-d) Comparison of our regression models MLP-Embd, GRU, and GRU-Embd trained on the entire CHANGE-seq and GUIDE-seq data through transfer learning with state-of-the-art-models CRISPR-Net and CRISPR-IP on our refined TrueOT benchmark dataset after removing experiments on which CRISPR-Net was trained (a,c) and its only-bulges subset (b,d).

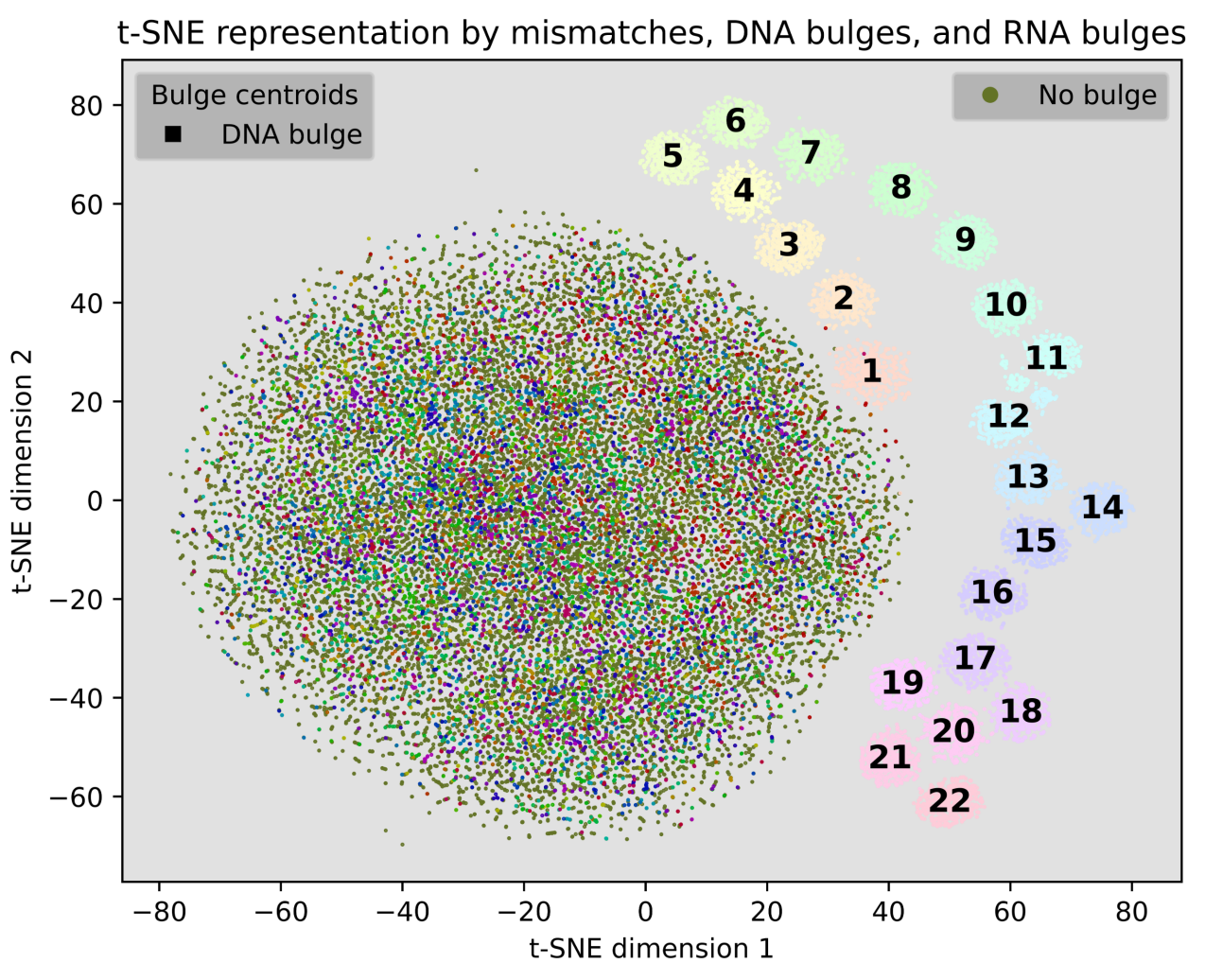

Figure S3. Visualization of the one-hot-encoding sequence features of off-target sites with bulges. One-hot-encoding sequence features for a pair of aligned OTS and their sgRNAs projected into a two-dimensional map using t-SNE. Comparing OTS with mismatches to OTS with bulges according to their bulge type and location. The centroid of OTS with DNA bulge location is illustrated in each map by the position of the label. We did not illustrate the centroids of OTS with RNA bulge as they are too dispersed in the visual representation.

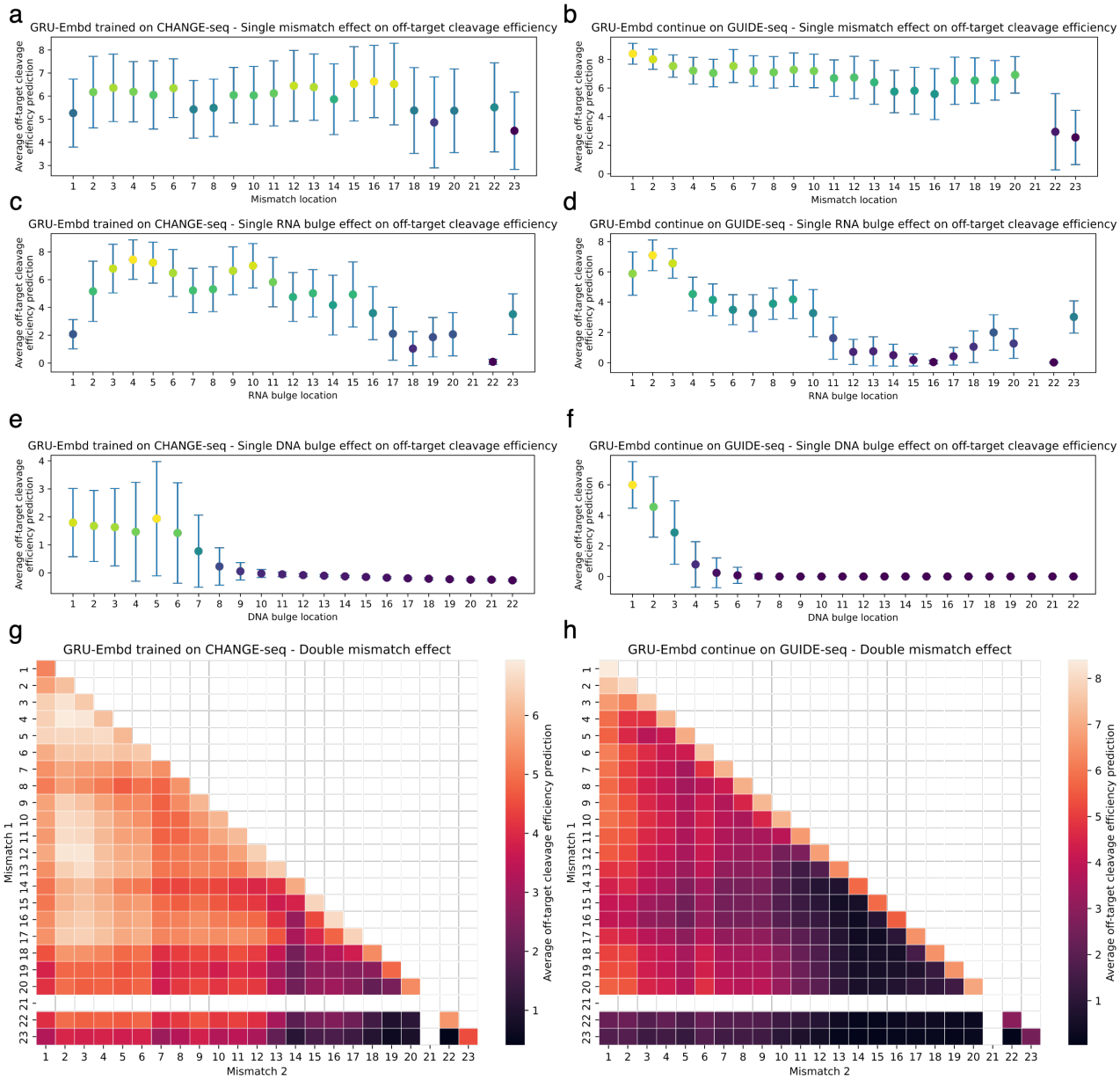

Figure S4. Positional effect of mismatches and bulges on off-target activity learned by our GRU-Embd model. a-f Positional effect of a single mismatch (a-b), RNA bulge (c-d), or DNA bulge (e-f) compared to the sgRNA predicted by a model for *in vitro* (a, c, e) and *in cellula* (b, d, f) data. **g-h** Positional effect of two mismatches compared to the sgRNA predicted by a model for *in vitro* (g) and *in cellula* (h) data.