ACS Synthetic Biology

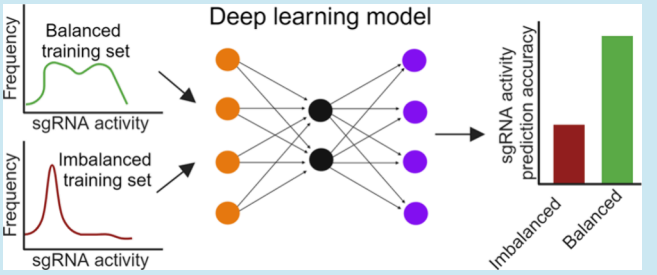
pubs.acs.org/synthbio
Technical Note

Balanced Training Sets Improve Deep Learning-Based Prediction of CRISPR sgRNA Activity

Varun Trivedi, Amirsadra Mohseni, Stefano Lonardi, and Ian Wheeldon*

Cite This: https://doi.org/10.1021/acssynbio.4c00542

ABSTRACT: CRISPR-Cas systems have transformed the field of synthetic biology by providing a versatile method for genome editing. The efficiency of CRISPR systems is largely dependent on the sequence of the constituent sgRNA, necessitating the development of computational methods for designing active sgRNAs. While deep learning-based models have shown promise in predicting sgRNA activity, the accuracy of prediction is primarily governed by the data set used in model training. Here, we trained a convolutional neural network (CNN) model and a large language model (LLM) on balanced and imbalanced data sets generated from CRISPR-Cas12a screening data for the yeast *Yarrowia lipolytica* and evaluated their ability to predict high- and low-activity sgRNAs. We further tested whether prediction performance can be improved by training on imbalanced data sets augmented with synthetic sgRNAs. Lastly, we demonstrated that adding synthetic sgRNAs to inherently imbalanced CRISPR-Cas9 data sets from *Y. lipolytica* and *Komagataella phaffii* leads to improved performance in predicting sgRNA activity, thus underscoring the importance of employing balanced training sets for accurate sgRNA activity prediction.



KEYWORDS: *sgRNA activity prediction, balanced training data sets, training set composition, deep learning, CRISPR genome editing*

INTRODUCTION

CRISPR systems are a potent tool for targeted genome editing in assays ranging from individual genetic perturbation experiments to high-throughput functional genetic screens.[1−4] CRISPR systems achieve efficient targeted editing by utilizing two components, a Cas endonuclease that creates a double-stranded break and a single guide RNA (sgRNA) that guides the Cas enzyme to the targeted genomic locus.[5,6] Genome editing efficacy depends on several factors such as the sequence and nucleotide composition of the sgRNA, propensity of the sgRNA to form secondary structure, genomic context, and epigenetic features like chromatin accessibility and DNA methylation.[7−10] As a result, CRISPR systems often have a broad spectrum of activity, with only a limited fraction of sgRNA successfully generating a desired genetic manipulation, thus emphasizing the need for computational approaches to design sgRNAs.

A host of computational tools for CRISPR sgRNA design have been developed that possess the ability to predict sgRNA activity in prokaryotic and eukaryotic organisms using machine learning and deep learning approaches.[11−14] These methods use large data sets that link sgRNA sequence with Cas activity as training sets to capture generalizable patterns and features of sgRNAs, and in doing so generate design rules for maximizing sgRNA activity.[15,16] The composition of the training data sets used as input to these methods thus plays a critical role in determining the accuracy of activity predictions. Training data sets consisting of a large number of sgRNAs with a wide distribution of activity lead to more accurate predictions of activity compared to skewed data sets.[12]

In this work, we evaluated the effect of training set composition on the performance of deep learning methods for sgRNA activity prediction. We trained a deep CNN model, DeepGuide,[12] and an LLM architecture, HyenaDNA,[18] on previously reported CRISPR-Cas12a data from *Y. lipolytica*,[12,17] and found the activity prediction accuracy with training sets skewed toward high- and low-activity sgRNAs to be lower relative to that with a balanced training set. Upon augmenting imbalanced training data sets with synthetic sgRNAs and retraining the models, we observed a partial recovery in predictive power lost during training on imbalanced data sets.

RESULTS AND DISCUSSION

Balanced Training Sets Enable Accurate Predictions of sgRNA Activity. The CRISPR-Cas12a sgRNA data set in *Y. lipolytica*, previously reported in refs 12 and 17, is an 8-fold coverage library containing 57018 sgRNA targeting ∼98% of

Received: August 9, 2024
Revised: October 9, 2024
Accepted: October 21, 2024

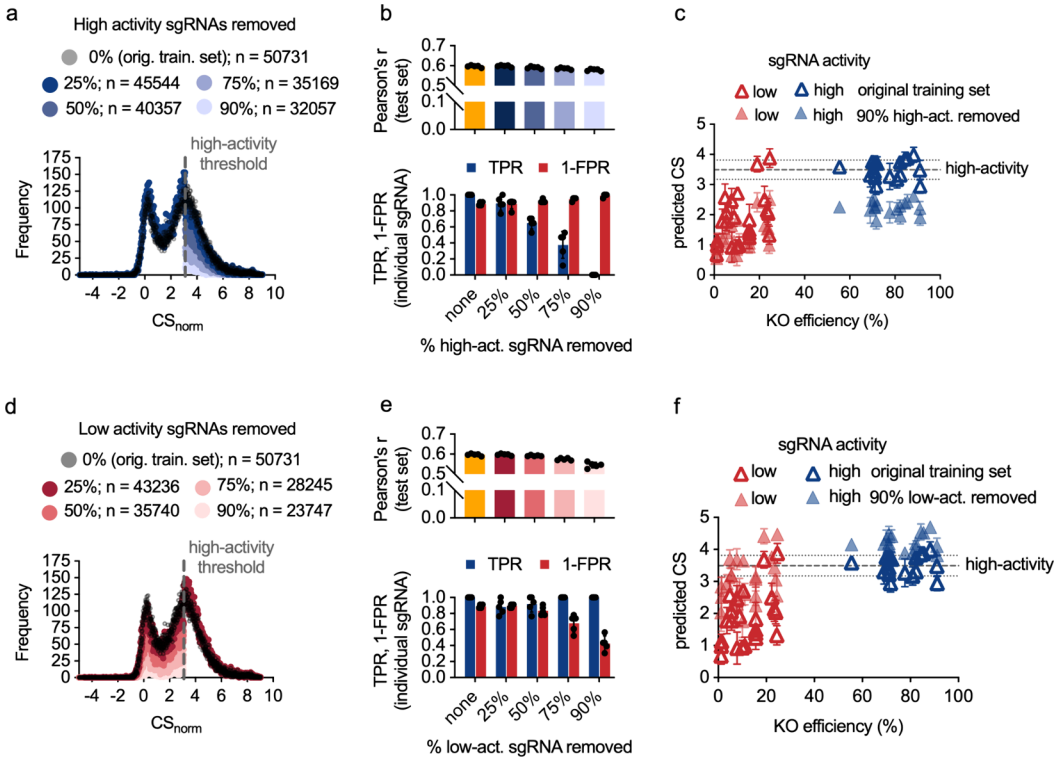


Figure 1. DeepGuide performance with imbalanced CRISPR-Cas12a training data sets. (a, d) Normalized cutting score (CS_{norm}) distributions of the Cas12a training data set imbalanced by removing 0, 25, 50, 75, and 90% high- and low-activity sgRNAs along with the total number of sgRNAs (n) in every training set. (b, e) Performance of DeepGuide models on the sgRNA test sets (Pearson's r), and high- and low-activity Cas12a sgRNAs from individual phenotype screening experiments (TPR, 1-FPR). Bars represent mean values of Pearson's r, TPR, and 1-FPR across five independent runs (n = 5). Error bars indicate one standard deviation, and data points represent values from each individual run. (c, f) Mean predicted CS of high-activity (KO efficiency ≥ 50%) and low-activity (KO efficiency < 50%) sgRNAs from individual phenotype screening experiments in *Y. lipolytica* when DeepGuide was trained on imbalanced data sets with 90% high-activity and 90% low-activity sgRNA removed. Data points represent mean values of predicted CS for experimental sgRNA with a given KO efficiency across five independent runs (n = 5), and error bars indicate one standard deviation. Dashed line represents average predicted CS threshold for high-activity, and dotted lines represent one standard deviation of the high-activity threshold.

the protein-coding genes in the PO 1f strain. Unbiased design and screening of this library produced a data set containing a well-balanced representation of high- and low-activity sgRNAs. The activity of each guide was determined using an experimental cutting score (CS), computed as the $\log_2$ ratio of sgRNA abundance in a strain deficient in nonhomologous end joining (NHEJ) to that in a Cas12a-expressing strain deficient in NHEJ.[12,19] Using this CS data set, we trained DeepGuide[12] to predict CRISPR-Cas12a activity based on sgRNA sequence. The data was split into training and test sets in the ratio 90:10, with the training set consisting of 50731 sgRNAs, comprising ∼41% high-activity sgRNA and ∼59% low-activity sgRNA (Figure S1). Model training resulted in a mean Pearson's r of 0.596 (experimental CS vs DeepGuide-predicted CS), thus, establishing a baseline of model performance for sgRNA activity predictions.

Imbalanced Training Sets Result in Poor Prediction of sgRNA Activity. In this experiment, we sought to quantify the impact of a skewed training set on model performance by randomly removing or adding high- or low-activity sgRNAs from a balanced training set and evaluating the performance of DeepGuide when trained on the imbalanced data sets. While Pearson's r computed for the test set captures the overall accuracy of the model in predicting the CS of each sgRNA, it does not gauge the model's ability to correctly predict high- and low-activity sgRNAs as measured in experimental assays.[20−22] For this reason, we evaluated the performance of DeepGuide to accurately classify a set of experimentally validated high- and low-activity sgRNAs identified from individual phenotype screening experiments.[12]

We first generated imbalanced training data sets biased toward low-activity sgRNAs by removing 25, 50, 75, and 90% high-activity sgRNAs from the original training set (Figure 1a). DeepGuide's performance on the test set was found to decrease slightly as more high-activity sgRNAs were removed (Figure 1b). Given that the Pearson's r for a balanced training set of 30000 sgRNA was found to be higher (r = 0.588; Figure S2) than that when 90% high-activity sgRNA were removed (r = 0.579; training set size ∼ 32000), the drop in Pearson's r can be attributed to a decrease in the share of high-activity sgRNAs in the training set. The True Positive Rate (TPR) was found to decline sharply as the percentage of high-activity sgRNAs removed increased, indicating a reduced ability to predict high-activity sgRNAs when the data set is skewed toward low-activity guides (Figure 1b). The decrease in TPR was accompanied by increases in 1 - False Positive Rate (1-FPR). As the training data sets become more biased toward low-activity sgRNAs, the DeepGuide-predicted CS of experimental sgRNAs shift to lower values leading to fewer sgRNAs being predicted as high-activity (Figures 1c and S3a,b).

Similar to the results with data sets biased toward low-activity sgRNAs, DeepGuide's performance decreased as the

https://doi.org/10.1021/acssynbio.4c00542
ACS Synth. Biol. XXXX, XXX, XXX−XXX

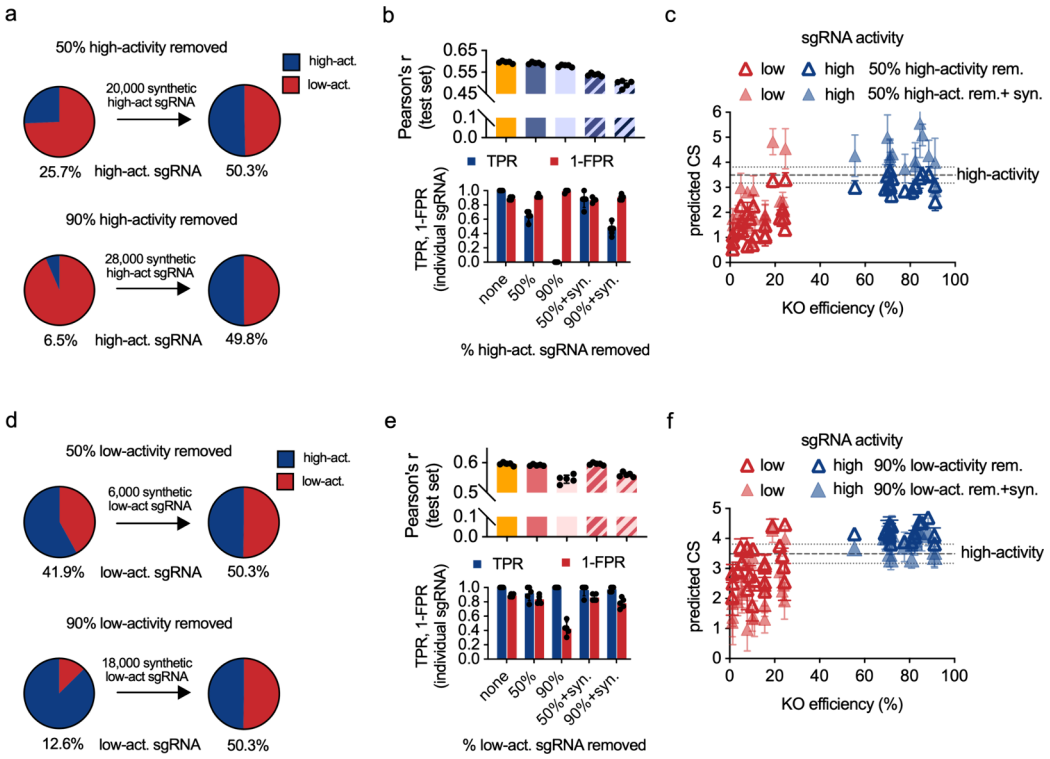


Figure 2. DeepGuide performance with imbalanced CRISPR-Cas12a training data sets augmented with synthetic sgRNAs. (a, d) Pie charts showing change in composition of imbalanced training sets skewed toward low- and high-activity sgRNAs after adding synthetic (a) high-activity and (d) low-activity sgRNAs. (b, e) Performance of DeepGuide models on the test set of sgRNAs (Pearson's *r*), and high-activity and low-activity Cas12a sgRNAs from individual phenotype screening experiments (TPR, 1-FPR), when trained using the original training set, imbalanced training sets obtained after removing 50% and 90% (b) high- and (e) low-activity sgRNAs, and rebalanced training sets obtained after adding synthetic (b) high- and (e) low-activity sgRNAs. Bars represent mean values of Pearson's *r*, TPR and 1-FPR across five independent runs ($n = 5$), error bars indicate one standard deviation, and data points represent values from each individual run. (c, f) Mean predicted CS of high-activity (KO efficiency $\geq$ 50%) and low-activity (KO efficiency < 50%) sgRNA from individual phenotype screening experiments in *Y. lipolytica* when DeepGuide was trained on balanced data sets containing synthetic (c) high- and (f) low-activity sgRNAs, with respect to the mean predicted CS of the same guides obtained upon training DeepGuide on the corresponding imbalanced data sets with 50% high-activity and 90% low-activity sgRNA removed. Data points represent mean values of predicted CS for experimental sgRNA with a given KO efficiency across five independent runs ($n = 5$), and error bars indicate one standard deviation. Dashed line represents average predicted CS threshold for high-activity and dotted lines represent one standard deviation of the high-activity threshold.

population was skewed toward high-activity (Figure 1d,e). As fewer low-activity sgRNA were retained in the training set, DeepGuide gradually lost the ability to accurately predict low-activity experimental data (Figure 1e). Figures 1f and S3a,c show that the predicted CS of experimental sgRNAs shift to higher values with respect to the predicted CS obtained by training DeepGuide on the original data set, ultimately causing fewer sgRNA to be predicted as low-activity.

We also generated imbalanced data sets by adding guides to a balanced data set containing 50% high- and low-activity sgRNA from the original training set. To generate data sets biased toward low-activity sgRNA, we added 25% (7495) and 50% (14991) low-activity sgRNA to the balanced training set. DeepGuide training on these data sets resulted in a slight increase in performance on the test set as more low-activity guides were added, likely due to an increase in training set size (Figure S4a). The TPR, however, was found to decrease as the fraction of high-activity sgRNA in the training data decreased, accompanied by small increases in 1-FPR, a result similar to that obtained with the previous method of creating imbalance by removing guides (Figure S4a). We also generated data sets skewed toward high-activity sgRNA by adding 25% (5187) and 50% (10374) high-activity sgRNA to the balanced training set. As expected, the decrease in proportion of low-activity sgRNA resulted in a decrease in 1-FPR, indicating reduced ability of the model to predict low-activity guides (Figure S4b).

To test whether other model frameworks exhibit similar behavior in predicting sgRNA activity when trained on imbalanced data sets, we evaluated the performance of HyenaDNA,[18] a large language model (LLM), trained on various data sets. For this analysis, we used the balanced training data set and the skewed data sets with 50% and 90% high- and low-activity sgRNAs removed. Similar to the DeepGuide results, Pearson's *r* slightly decreased as more high- and low-activity sgRNAs were removed (Figure S5a). Likewise, the TPR and 1-FPR decreased when the data sets were skewed toward low- and high-activity guides, respectively (Figure S5b). These results substantiate the importance of the training set composition in influencing sgRNA activity prediction accuracy, independent of the prediction model.

Augmenting Imbalanced Training Sets with Synthetic sgRNAs Helps Recover Activity Prediction Power. To examine whether artificially rebalancing training sets improves prediction performance, we augmented imbalanced CRISPR-Cas12a training sets with synthetic high- or low-activity sgRNA. CRISPR-Cas activity is less sensitive to mismatches between the sgRNA and target DNA in the PAM-distal region of the sgRNA compared to the PAM-

https://doi.org/10.1021/acssynbio.4c00542
ACS Synth. Biol. XXXX, XXX, XXX−XXX

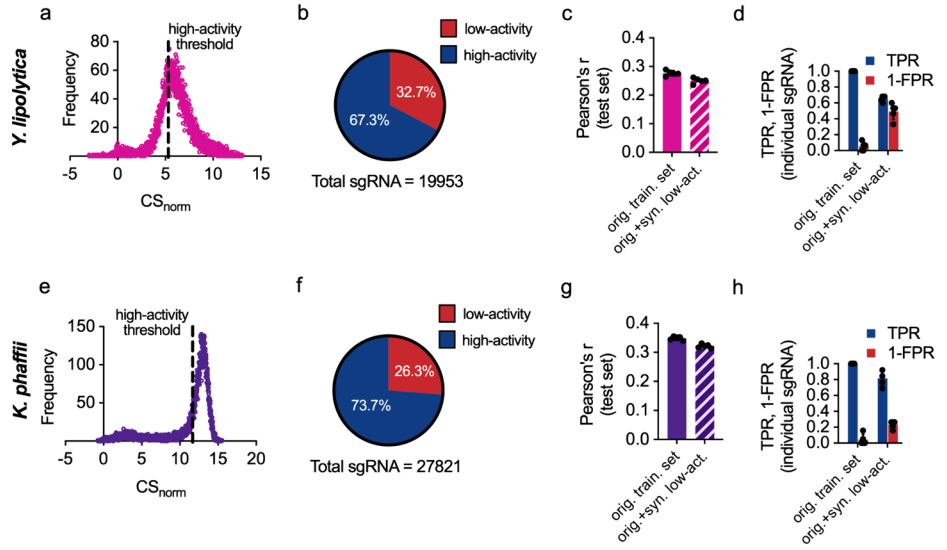


Figure 3. Composition of the *Y. lipolytica* (top) and *K. phaffii* (bottom) CRISPR-Cas9 training sets and DeepGuide performance with the two data sets. (a, e) Normalized cutting score (CS_{norm}) distributions of the original Cas9 training data sets for *Y. lipolytica* and *K. phaffii*. (b, f) Pie charts showing the proportion of high- and low-activity sgRNAs in the original Cas9 training sets containing a total of 19953 sgRNA for *Y. lipolytica* and 27821 sgRNA for *K. phaffii*. (c, g) Performance of DeepGuide on the test set of sgRNA for *Y. lipolytica* and *K. phaffii* when trained on the respective original training sets and rebalanced training sets obtained after adding synthetic low-activity sgRNA to the original sets. Bars represent mean Pearson's *r* across five independent runs ($n = 5$), error bars indicate one standard deviation, and data points represent values from each individual run. (d, h) DeepGuide performance on high- and low-activity Cas9 sgRNAs from individual experiments for *Y. lipolytica* and *K. phaffii*. Bars represent mean values of TPR and 1-FPR across five independent runs ($n = 5$), error bars indicate one standard deviation, and data points represent values from each individual run.

proximal or seed region.[23−25] For CRISPR-Cas12a, the first 14 bp of a sgRNA from the 5′ end comprise the seed region.[23,26] We thus generated synthetic sgRNAs by randomly sampling guides from the minority class in a given imbalanced training set (for a training set biased toward low-activity sgRNA, the minority class constitutes all high-activity sgRNAs within the training set, and vice versa) and created new guides with random one nucleotide substitution in the nonseed region (base positions 15−25 from the 5′ end) of the selected guide. Since the CRISPR-Cas12a library was designed by ensuring the uniqueness of the 14 bp sgRNA seed region in the genome,[12,17] the generated synthetic sgRNA would always target the same genomic locus as the original sgRNA it was created from.

To rebalance training sets biased toward low-activity sgRNA, we augmented the data sets consisting of 50% and 90% high-activity sgRNAs removed with 20000 and 28000 synthetic high-activity sgRNA, respectively (Figure 2a), that were generated by sampling high-activity sgRNA from the corresponding imbalanced data sets. DeepGuide training on these rebalanced data sets resulted in a small decrease in performance (Pearson's *r*) on the test set compared to that for the corresponding imbalanced training sets (Figure 2b). The TPR for experimental high-activity sgRNAs, however, exhibited an increase when synthetic high-activity sgRNAs were added to the training sets. For the data set with 50% high-activity sgRNA removed, the recovery in performance yields predictions that closely match those achieved using the original training set (Figure 2b). The 1-FPR value, meanwhile, showed a small decrease when synthetic high-activity sgRNA were appended to the training sets, while still remaining above 0.85 for all data sets. Figure 2c shows that the addition of synthetic high-activity sgRNAs to imbalanced training sets causes the predicted CS of experimental sgRNAs to shift to higher values, illustrating the recovery in the high-activity guide prediction accuracy. The data set containing 50% low-activity sgRNA added was augmented with 20000 synthetic high-activity sgRNA (resulting in a data set containing 50.3% high-activity sgRNA and 49.7% low-activity sgRNA), and once again, training on this rebalanced data set led to an increase in TPR compared to the imbalanced data set (Figure S6a).

We next rebalanced training data sets with 50% and 90% low-activity sgRNAs removed by augmenting them with 6000 and 18000 synthetic low-activity sgRNAs, respectively (Figure 2d). We also augmented the data set containing 50% high-activity sgRNA added with 6000 synthetic low-activity sgRNA (resulting in a data set containing 50.3% low-activity sgRNA and 49.7% high-activity sgRNA). This resulted in minimal change in Pearson's *r* for DeepGuide predictions on the test set (Figures 2e and S6b). More importantly, the addition of synthetic sgRNAs led to an increase in 1-FPR for both data sets. It is noteworthy here that model performance for the data set supplemented with synthetic sgRNAs after removing 90% of the low-activity population is inferior to that for the original data set only by a small margin. Figure 2f illustrates the shift in predicted CS of experimental sgRNAs to lower values upon addition of synthetic low-activity sgRNA to imbalanced training data sets.

We also explored variations of the approach to generate synthetic sgRNA, and we investigated their ability to improve prediction performance. Addition of synthetic sgRNA generated using the different methods (double mutants, CS penalty, and others; see Methods) resulted in a similar performance on the test set and were not an improvement over the method shown in Figure 2 (Figure S7). Overall, the similar performance of variant methods implies that the method used for generating synthetic sgRNA has no effect on the improvement in model performance.

https://doi.org/10.1021/acssynbio.4c00542
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Adding Synthetic sgRNA to Imbalanced CRISPR-Cas9 Data Sets Improves Low-Activity sgRNA Prediction. To assess the capability of the synthetic sgRNA-based approach in improving activity prediction on imbalanced training sets from other species and endonucleases, we implemented DeepGuide on CRISPR-Cas9 data sets from *Y. lipolytica* and *K. phaffii* previously reported in refs 27 and 28. The *Y. lipolytica* Cas9 data set is biased toward high-activity sgRNA; the set includes 67.3% high-activity sgRNA with a training set size of 19953 (Figure 3a,b). To alleviate this imbalance, we augmented the training set with 6500 synthetic low-activity sgRNA by creating a 1 bp substitution in the nonseed region. Since DeepGuide improves Cas9 activity predictions using nucleosome occupancy information,[12] we provided occupancy scores for every sgRNA in addition to sgRNA sequence as input for training on the Cas9 data sets. When trained on the original and rebalanced training sets, DeepGuide was found to yield nearly similar values of Pearson's *r* on the test set of sgRNA (Figure 3c). Addition of synthetic sgRNA also resulted in an increase in 1-FPR from 0.053 to 0.493 for the original and rebalanced data sets, respectively, but at the cost of a decrease in TPR from 1 to 0.656, Figure 3d. Figure S8a shows the predicted CS of experimental high- and low-activity sgRNA before and after the addition of synthetic sgRNA to the original training set.

The *K. phaffii* training set contains a disproportionately large number of high-activity sgRNA (73.7% high-activity sgRNA in a training set of 27821 sgRNA, Figure 3e,f) and was, hence, rebalanced by adding 13000 synthetic low-activity sgRNA. DeepGuide implementation on the training sets resulted in similar values of Pearson's *r* (Figure 3g). More prominently, when measuring performance on experimentally validated high- and low-activity sgRNA from individual experiments,[28] the addition of synthetic low-activity sgRNA led to an jump in 1-FPR from 0.042 to 0.232, accompanied by a small decrease in TPR from 1 to 0.815 (Figures 3h and S8b).

Deep learning models, while having been shown to be effective in designing sgRNAs, depend significantly on the training set composition for accurate prediction of activity. Implementation of deep learning models on CRISPR-Cas data sets in this study shows that adding synthetic sgRNAs can improve performance with imbalanced data sets, but not to the level of balanced data sets. Ultimately, AI models result in best prediction performance when trained on data sets evenly representing both positive and negative biological outcomes, or well-balanced data sets.

METHODS

Processing *Y. lipolytica* and *K. phaffii* CRISPR-Cas12a and CRISPR-Cas9 sgRNA-CS Data. *Y. lipolytica* sgRNA sequence and CS data for the CRISPR-Cas12a library was obtained from ref 12, while CRISPR-Cas9 data sets for *Y. lipolytica* and *K. phaffii* were obtained from refs 27 and 28, respectively. For all data sets, raw CS values of sgRNA were converted to normalized CS by subtracting the average CS of all nontargeting sgRNA in the respective libraries from the raw CS values of every sgRNA. For Cas12a data, the 25 bp sequences of sgRNA were extended to 32 bp sequences (25 bp spacer +4 bp PAM + 1 bp context upstream of the PAM + 2 bp context downstream of the spacer) using custom Python scripts to map sgRNA to the *Y. lipolytica* CLIB89 genome (https://www.ncbi.nlm.nih.gov/assembly/GCA_001761485.1)[29] and obtain the upstream and downstream nucleotides. In case of the Cas9 data sets, the 20 bp sequences of sgRNA were extended to 28 bp sequences (20 bp spacer +3 bp PAM + 2 bp context upstream of the spacer +3 bp context downstream of the PAM) by mapping sgRNA to *Y. lipolytica* CLIB89 and *K. phaffii* GS115 (https://www.ncbi.nlm.nih.gov/assembly/GCA_000027005.1)[30] genomes.

For each data set, the "sgRNA + PAM + upstream/downstream context" sequences and normalized CS data were then randomly split into training and test sets for the sgRNA activity prediction tools in the ratio 90:10. For *Y. lipolytica* CRISPR-Cas12a data, the original training set consisted of 50731 sgRNA, while the test set comprised 5637 sgRNA. Guides in the original training set were classified as high-activity and low-activity based on a high-activity threshold defined in ref 17, equivalent to a normalized CS of 3.10. The training and test sets for *Y. lipolytica* CRISPR-Cas9 data consisted of 19953 and 2217 sgRNA, respectively, with a high-activity threshold equivalent to normalized CS of 5.30, as defined in ref 27. Similarly, for *K. phaffii* Cas9 data, the training and test set sizes were 27821 and 3093 sgRNA, respectively, with sgRNA having normalized CS greater than 11.66 deemed as high-activity sgRNA, based on the threshold defined in ref 28.

DeepGuide Implementation. For *Y. lipolytica* data sets, DeepGuide (https://github.com/ucrbioinfo/deepguide_reborn)[12] was first pretrained on the *Y. lipolytica* CLIB89 genome using a sequence length of 32 bp for Cas12a (guide_length: 32) and 28 bp (guide_length: 28) for Cas9 with 6 epochs (dg_one_pretrain_epochs: 6), followed by training on the *Y. lipolytica* Cas12a/Cas9 data with 10 epochs (dg_one_epochs: 10). For the *K. phaffii* Cas9 data set, the pretraining was performed on the *K. phaffii* GS115 genome using 28 bp as the sequence length (guide_length: 28).

Both the pretraining and training steps were performed using a batch size of 64 (dg_one_pretrain_batch_size: 64, and dg_one_batch_size: 64) and a train/validation split of 70:30 (dg_one_pretrain_train_test_ratio: 0.7, and dg_one_train_test_ratio: 0.7). For Cas12a, the "cas" parameter was set to 'cas9_seq', since only sgRNA sequence data was used for training. For Cas9 data sets, the value of the "cas" parameter was changed to 'cas9_nucleosome', since sgRNA nucleosome occupancy scores were used for training in addition to sequence data. Five independent runs were performed for each experiment.

HyenaDNA Implementation. HyenaDNA (https://github.com/HazyResearch/hyena-dna)[18] was pretrained on the *Y. lipolytica* CLIB89 genome using a sequence length of 32 bp (max_length: 32), train/val/test split of 80:10:10, model width of 32 (d_model: 32), depth of 2 layers (n_layer: 2), a learning rate of 6×10^{-4} (lr: 6e-4), and a global batch size of 1024 (global_batch_size: 1024) with 100 epochs (max_epochs: 100). Default values of all other parameters were used. Pretraining was carried out on 4 Nvidia A100 80GB GPUs (devices: 4).

For fine-tuning the model, the Cas12a training data was split into training and validation sets in the ratio 80:10 (train_len: 45094 for the original training set), and a global batch size of 256 (global_batch_size: 256) was used. The model configuration, sequence length, and learning rate were kept unchanged from the pretraining step (d_model: 32, n_layer: 2, max_length: 32, lr: 6e-4). The fine-tuning step was also performed with 100 epochs (max_epochs: 100) using one Nvidia A100 80GB GPU (devices: 1), and the entire model was fine-tuned rather than freezing the weights of the

https://doi.org/10.1021/acssynbio.4c00542
ACS Synth. Biol. XXXX, XXX, XXX−XXX

pretrained backbone (freeze_backbone: false). Default values of all other parameters were used. Five independent runs were performed for each experiment.

Generating Imbalanced Data Sets by Adding High/Low-Activity sgRNA. A reduced balanced data set containing 50% high-activity and low-activity sgRNA from the original Cas12a training set (50731 sgRNA) was first created. This data set thus contains 25366 sgRNA, constituting 10375 high-activity sgRNA and 14991 low-activity sgRNA. To skew this balanced data set toward low-activity sgRNA, 25% (7495) and 50% (14991) low-activity sgRNA were added by sampling from the pool of sgRNA not present in the balanced training set, resulting in data sets containing 75% and 100% of all low-activity training guides, respectively. Similarly, the balanced training set was skewed toward high-activity sgRNA by adding 25% (5187) and 50% (10374) high-activity sgRNA to generate data sets containing 75% and 100% of all high-activity training guides, respectively.

Generating Synthetic sgRNA. Custom Python scripts were used to generate synthetic sgRNA by randomly sampling appropriate number of sgRNA from the pool of high-/low-activity sgRNA in the imbalanced training sets, and creating a 1 bp substitution for four of the five simulation methods, and 2 bp substitutions for one method, in the nonseed region (base positions 15−25 from the 5′ end on the 25 bp spacer sequence for Cas12a sgRNA and positions 1−8 from the 5′ end of the 20 bp spacer for Cas9 sgRNA[31,32]) of the sampled sgRNA.

In the case of unbiased sampling with penalized CS for Cas12a sgRNA, the normalized CS of the synthetic guides was reduced by 1/25th, or 4% compared to that of the original sgRNA to account for a possible reduction in sgRNA activity due to a 1 bp mismatch.

For biased sampling toward sgRNA with extremely high/low CS values, positive and negative exponential distributions were created for the range of normalized CS values for high-activity and low-activity guides, respectively. For every simulated guide, a random value was sampled from this exponential distribution, and the normalized CS value closest to this sampled value and the corresponding sgRNA sequence were used to generate the synthetic guide.

For creating substitutions by biasing toward terminal positions on the sgRNA, the position for creating a substitution was sampled from an exponential distribution so that the probability of sampling terminal positions is higher compared to relatively central positions.

Computing Nucleosome Occupancy Scores. Genome-wide nucleosome occupancy data for *Y. lipolytica* CLIB89 and *K. phaffii* GS115 genomes were obtained from MNase-seq data sets previously reported in refs 33 and 34, respectively. For every Cas9 sgRNA, an average occupancy score of the corresponding target locus was first computed by averaging the occupancy scores across all target bases, followed by normalizing the scores to values between 0 and 1 by dividing each average score by the highest average score in the respective data set (*Y. lipolytica*/*K. phaffii*). The average normalized occupancy scores obtained for each sgRNA were then used to train DeepGuide alongside the sgRNA sequence information.

Calculation of TPR and 1-FPR. Based on the predicted CS of sgRNA from individual phenotype screening experiments, every sgRNA was classified as high-activity or low-activity, which was different from the high/low-activity classification based on experimental knockout efficiency. The predicted high/low-activity classification was based on a *p*-value derived from a *z*-test of significance. Briefly, predicted CS values of experimental high-activity sgRNA (i.e., sgRNA having knockout efficiency ≥ 50%) obtained from the models trained on the original training sets were used to create a population of predicted CS of high-activity sgRNA for the respective data sets. Predicted CS values of experimental low-activity sgRNA obtained from the models trained on the original training sets, as well as predicted CS values of all (i.e., experimental high-activity and low-activity) sgRNA in every subsequent activity prediction trial were compared to this population in a *z*-test of significance to determine if a given predicted CS value belongs to this population ($p > 0.05$; predicted high-activity sgRNA) or is significantly different from the population ($p < 0.05$; predicted low-activity sgRNA). The ability of a model to accurately predict sgRNA from individual experiments as high-activity and low-activity was measured using two metrics, True Positive Rate (TPR) and 1-False Positive Rate (FPR), respectively. TPR is defined as

$$\text{TPR} = \left(\frac{\text{No. of exptl high-activity sgRNA predicted to have high activity}}{\text{total No. of exptl high-activity sgRNA}}\right) \quad (1)$$

Similarly, 1-FPR is calculated as

$$\text{1-FPR} = \left(\frac{\text{No. of exptl low-activity sgRNA predicted to have low activity}}{\text{total No. of exptl low-activity sgRNA}}\right) \quad (2)$$

Since the predicted CS values of experimental high-activity sgRNA from the model trained on the original set were used to generate the predicted high-activity population, all of these sgRNA were deemed to have high predicted activity, resulting in a TPR of 1 for the original training sets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.4c00542.

Additional figures pertaining to the deep learning experiments performed using balanced and imbalanced CRISPR-Cas12a/Cas9 data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ian Wheeldon − *Department of Chemical and Environmental Engineering, Integrative Institute for Genome Biology, and Center for Industrial Biotechnology, University of California, Riverside, California 92521, United States;* orcid.org/0000-0002-3492-7539; Email: wheeldon@ucr.edu

Authors

Varun Trivedi − *Department of Chemical and Environmental Engineering, University of California, Riverside, California 92521, United States*

Amirsadra Mohseni − *Department of Computer Science, University of California, Riverside, California 92521, United States*

Stefano Lonardi − *Department of Computer Science and Integrative Institute for Genome Biology, University of California, Riverside, California 92521, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.4c00542

https://doi.org/10.1021/acssynbio.4c00542
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Author Contributions

V.T., S.L., and I.W. conceived the project and planned the experiments. V.T. and A.M. performed DeepGuide experiments with *Y. lipolytica* and *K. phaffii* data sets. V.T. performed LLM experiments with *Y. lipolytica* CRISPR-Cas12a data. All authors wrote and edited the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by NSF CBET 2225878.

■ ABBREVIATIONS

CNN - convolutional neural network
LLM - large language model
CS - cutting score
NHEJ - nonhomologous end joining
TPR - true positive rate
FPR - false positive rate

■ REFERENCES

(1) Przybyla, L.; Gilbert, L. A. A New Era in Functional Genomics Screens. *Nat. Rev. Genet.* **2022**, *23* (2), 89−103.
(2) Shalem, O.; Sanjana, N. E.; Zhang, F. High-Throughput Functional Genomics Using CRISPR−Cas9. *Nat. Rev. Genet.* **2015**, *16* (5), 299−311.
(3) Hart, T.; Tong, A. H. Y.; Chan, K.; Van Leeuwen, J.; Seetharaman, A.; Aregger, M.; Chandrashekhar, M.; Hustedt, N.; Seth, S.; Noonan, A.; Habsid, A.; Sizova, O.; Nedyalkova, L.; Climie, R.; Tworzyanski, L.; Lawson, K.; Sartori, M. A.; Alibeh, S.; Tieu, D.; Masud, S.; Mero, P.; Weiss, A.; Brown, K. R.; Usaj, M.; Billmann, M.; Rahman, M.; Costanzo, M.; Myers, C. L.; Andrews, B. J.; Boone, C.; Durocher, D.; Moffat, J. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Genes|Genomes|Genetics* **2017**, *7* (8), 2719−2727.
(4) Doench, J. G. Am I Ready for CRISPR? A User's Guide to Genetic Screens. *Nat. Rev. Genet.* **2018**, *19* (2), 67−80.
(5) Doudna, J. A.; Charpentier, E. The New Frontier of Genome Engineering with CRISPR-Cas9. *Science* **2014**. DOI: 10.1126/science.1258096.
(6) Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; Charpentier, E. A Programmable Dual-RNA−Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **2012**. *337*816.
(7) Javaid, N.; Choi, S. CRISPR/Cas System and Factors Affecting Its Precision and Efficiency. *Front. Cell Dev. Biol.* **2021**, *9*, No. 761709.
(8) Horlbeck, M. A.; Witkowsky, L. B.; Guglielmi, B.; Replogle, J. M.; Gilbert, L. A.; Villalta, J. E.; Torigoe, S. E.; Tjian, R.; Weissman, J. S. Nucleosomes Impede Cas9 Access to DNA in Vivo and in Vitro. *eLife* **2016**, *5*, na.
(9) Yarrington, R. M.; Verma, S.; Schwartz, S.; Trautman, J. K.; Carroll, D. Nucleosomes Inhibit Target Cleavage by CRISPR-Cas9 in Vivo. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (38), 9351−9358.
(10) Přibylová, A.; Fischer, L.; Pyott, D. E.; Bassett, A.; Molnar, A. DNA Methylation Can Alter CRISPR/Cas9 Editing Frequency and DNA Repair Outcome in a Target-Specific Manner. *New Phytol.* **2022**, *235* (6), 2285−2299.
(11) Doench, J. G.; Hartenian, E.; Graham, D. B.; Tothova, Z.; Hegde, M.; Smith, I.; Sullender, M.; Ebert, B. L.; Xavier, R. J.; Root, D. E. Rational Design of Highly Active sgRNAs for CRISPR-Cas9−mediated Gene Inactivation. *Nat. Biotechnol.* **2014**, *32* (12), 1262−1267.
(12) Baisya, D.; Ramesh, A.; Schwartz, C.; Lonardi, S.; Wheeldon, I. Genome-Wide Functional Screens Enable the Prediction of High Activity CRISPR-Cas9 and -Cas12a Guides in Yarrowia Lipolytica. *Nat. Commun.* **2022**, *13* (1), 922.
(13) Guo, J.; Wang, T.; Guan, C.; Liu, B.; Luo, C.; Xie, Z.; Zhang, C.; Xing, X.-H. Improved sgRNA Design in Bacteria via Genome-Wide Activity Profiling. *Nucleic Acids Res.* **2018**, *46* (14), 7052−7069.
(14) Kim, H. K.; Min, S.; Song, M.; Jung, S.; Choi, J. W.; Kim, Y.; Lee, S.; Yoon, S.; Kim, H. Deep Learning Improves Prediction of CRISPR−Cpf1 Guide RNA Activity. *Nat. Biotechnol.* **2018**, *36* (3), 239−241.
(15) Sherkatghanad, Z.; Abdar, M.; Charlier, J.; Makarenkov, V. Using Traditional Machine Learning and Deep Learning Methods for on- and off-Target Prediction in CRISPR/Cas9: A Review. *Brief. Bioinform.* **2023**, *24* (3), No. bbad131.
(16) Konstantakos, V.; Nentidis, A.; Krithara, A.; Paliouras, G. CRISPR−Cas9 gRNA Efficiency Prediction: An Overview of Predictive Tools and the Role of Deep Learning. *Nucleic Acids Res.* **2022**, *50* (7), 3616−3637.
(17) Ramesh, A.; Trivedi, V.; Lee, S.; Tafrishi, A.; Schwartz, C.; Mohseni, A.; Li, M.; Lonardi, S.; Wheeldon, I. acCRISPR: An Activity-Correction Method for Improving the Accuracy of CRISPR Screens. *Commun. Biol.* **2023**, *6* (1), 617.
(18) Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; Ermon, S.; Baccus, S. A.; Ré, C. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *arXiv:2306.15794 [cs.LG]* **2023**, na.
(19) Schwartz, C.; et al. Validating Genome-Wide CRISPR-Cas9 Function Improves Screening in the Oleaginous Yeast Yarrowia Lipolytica. *Metab. Eng.* **2019**, *55*, 102−110.
(20) Asuero, A. G.; Sayago, A.; González, A. G. The Correlation Coefficient: An Overview. *Crit. Rev. Anal. Chem.* **2006**, *36*, 41.
(21) Reed, A. H. Misleading Correlations in Clinical Applications. *Clin. Chim. Acta* **1972**, *40* (1), 266−268.
(22) Westgard, J. O.; Hunt, M. R. Use and Interpretation of Common Statistical Tests in Method-Comparison Studies. *Clin. Chem.* **1973**, *19* (1), 49−57.
(23) Rabinowitz, R.; Offen, D. Single-Base Resolution: Increasing the Specificity of the CRISPR-Cas System in Gene Editing. *Mol. Ther.* **2021**, *29* (3), 937−948.
(24) Swarts, D. C.; van der Oost, J.; Jinek, M. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Mol. Cell* **2017**, *66* (2), 221−233.e4.
(25) Kim, H.; Lee, W.-J.; Oh, Y.; Kang, S.-H.; Hur, J. K.; Lee, H.; Song, W.; Lim, K.-S.; Park, Y.-H.; Song, B.-S.; Jin, Y. B.; Jun, B.-H.; Jung, C.; Lee, D.-S.; Kim, S.-U.; Lee, S. H. Enhancement of Target Specificity of CRISPR−Cas12a by Using a Chimeric DNA−RNA Guide. *Nucleic Acids Res.* **2020**, *48* (15), 8601−8616.
(26) Ramesh, A.; Wheeldon, I. Guide RNA Design for Genome-Wide CRISPR Screens in Yarrowia Lipolytica. *Yarrowia lipolytica* **2021**, *2307*, 123−137.
(27) Robertson, N. R.; Trivedi, V.; Lupish, B.; Ramesh, A.; Aguilar, Y.; Arteaga, A.; Nguyen, A.; Lee, S.; Lenert-Mondou, C.; Harland-Dunaway, M.; Jinkerson, R.; Wheeldon, I. Optimized Genome-Wide CRISPR Screening Enables Rapid Engineering of Growth-Based Phenotypes in Yarrowia Lipolytica. *Metab. Eng.* **2024**, 55−65.
(28) Tafrishi, A.; et al. Functional Genomic Screening in Komagataella Phaffii Enabled by High-Activity CRISPR-Cas9 Library. *Metab. Eng.* **2024**, *85*, 73−83.
(29) Magnan, C.; Yu, J.; Chang, I.; Jahn, E.; Kanomata, Y.; Wu, J.; Zeller, M.; Oakes, M.; Baldi, P.; Sandmeyer, S. Sequence Assembly of Yarrowia Lipolytica Strain W29/CLIB89 Shows Transposable Element Diversity. *PLoS One* **2016**, *11* (9), No. e0162363.
(30) De Schutter, K.; Lin, Y.-C.; Tiels, P.; Van Hecke, A.; Glinka, S.; Weber-Lehmann, J.; Rouzé, P.; Van de Peer, Y.; Callewaert, N. Genome Sequence of the Recombinant Protein Production Host Pichia Pastoris. *Nat. Biotechnol.* **2009**, *27* (6), 561−566.
(31) Hsu, P. D.; Scott, D. A.; Weinstein, J. A.; Ran, F. A.; Konermann, S.; Agarwala, V.; Li, Y.; Fine, E. J.; Wu, X.; Shalem, O.; Cradick, T. J.; Marraffini, L. A.; Bao, G.; Zhang, F. DNA Targeting Specificity of RNA-Guided Cas9 Nucleases. *Nat. Biotechnol.* **2013**, *31* (9), 827−832.

https://doi.org/10.1021/acssynbio.4c00542
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(32) Jiang, W.; Bikard, D.; Cox, D.; Zhang, F.; Marraffini, L. A. RNA-Guided Editing of Bacterial Genomes Using CRISPR-Cas Systems. *Nat. Biotechnol.* **2013**, *31* (3), 233−239.
(33) Tsankov, A. M.; Thompson, D. A.; Socha, A.; Regev, A.; Rando, O. J. The Role of Nucleosome Positioning in the Evolution of Gene Regulation. *PLoS Biol.* **2010**, *8* (7), No. e1000414.
(34) Liachko, I.; Youngblood, R. A.; Tsui, K.; Bubb, K. L.; Queitsch, C.; Raghuraman, M. K.; Nislow, C.; Brewer, B. J.; Dunham, M. J. GC-Rich DNA Elements Enable Replication Origin Activity in the Methylotrophic Yeast Pichia Pastoris. *PLoS Genet.* **2014**, *10* (3), No. e1004169.

https://doi.org/10.1021/acssynbio.4c00542
ACS Synth. Biol. XXXX, XXX, XXX−XXX