DeepCRISTL: deep transfer learning to predict CRISPR/Cas9 functional and endogenous on-target editing efficiency

Shai Elkayam and Yaron Orenstein

School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel

*To whom correspondence should be addressed.

Abstract

Motivation: CRISPR/Cas9 technology has been revolutionizing the field of gene editing in recent years. Guide RNAs (gRNAs) enable Cas9 proteins to target specific genomic loci for editing. However, editing efficiency varies between gRNAs. Thus, computational methods were developed to predict editing efficiency for any gRNA of interest. High-throughput datasets of Cas9 editing efficiencies were produced to train machine-learning models to predict editing efficiency. However, these high-throughput datasets have low correlation with functional and endogenous editing. Another difficulty arises from the fact that functional and endogenous editing efficiency is more difficult to measure, and as a result, functional and endogenous datasets are too small to train accurate machine-learning models on.

Results: We developed DeepCRISTL, a deep-learning model to predict the on-target efficiency given a gRNA sequence. DeepCRISTL takes advantage of high-throughput datasets to learn general patterns of gRNA on-target editing efficiency, and then uses transfer learning (TL) to fine-tune the model and fit it to the functional and endogenous prediction task. We pre-trained the DeepCRISTL model on more than 130 000 gRNAs, produced through the DeepHF study as a high-throughput dataset of three Cas9 enzymes. We improved the DeepHF model by multi-task and ensemble techniques and achieved state-of-the-art results over each of the three enzymes: up to 0.89 in Spearman correlation between predicted and measured on-target efficiencies. To fine-tune model weights to predict on-target efficiency of functional or endogenous datasets, we tested several TL approaches, with gradual learning being the overall best performer, both when pre-trained on DeepHF and when pre-trained on CRISPROn, another high-throughput dataset. DeepCRISTL outperformed state-of-the-art methods on all functional and endogenous datasets. Using saliency maps, we identified and compared the important features learned by the model in each dataset. We believe DeepCRISTL will improve prediction performance in many other CRISPR/Cas9 editing contexts by leveraging TL to utilize both high-throughput datasets, and smaller and more biologically relevant datasets, such as functional and endogenous datasets.

Availability and implementation: DeepCRISTL is available via github.com/OrensteinLab/DeepCRISTL.

Contact: yaronore@bgu.ac.il

Supplementary information: Supplementary data are available at Bioinformatics online.
hundreds). Hence, the main challenge arising from these datasets is how to utilize the high-throughput datasets for the task of functional or endogenous gene-editing prediction.

In recent years, deep neural networks have been revolutionizing the machine-learning field with the availability of abundant datasets and increased efficiency in computational power (Barshai et al., 2020). This revolution has sparked great interest and a body of applications in the bioinformatics domain. Deep neural networks were specifically successfully applied in predicting CRISPR/Cas9 on-target efficiencies (Wang et al., 2020). Seq-deepCpf1 is the first study to use deep learning to solve the challenge of predicting on-target efficiencies from CRISPR high-throughput data (Kim et al., 2018). In subsequent studies, Seq-deepCpf1 was outperformed by other deep-learning models, such as DeepHF (Wang et al., 2019) and CRISPRoN (Xiang et al., 2021). Even though those models were trained over large datasets of high-throughput data, and achieved superb performance in cross-validation (e.g. Spearman correlation of 0.87 on a held-out test subset of the DeepHF dataset), their performance on endogenous and functional datasets is much lower (e.g. Spearman correlation < 0.5; Wang et al., 2019).

One of the disadvantages of deep-learning models is their reliance on large datasets for accurate predictions (Barshai et al., 2020). Training models with many parameters, as in deep neural networks, may over-fit when trained on small datasets. One solution to the problem of training on small datasets is transfer learning (TL; Tan et al., 2018). In TL, a model is trained on a large dataset, which is referred to as the source data, and then fine-tuned on a small dataset, which is referred to as the target data. In order for TL to improve prediction performance, the source data has to correlate with the target data. Several TL approaches have been developed, where the most popular is the last-layer approach. In the last-layer approach, only the weights of the last hidden layer are re-trained in the fine-tuning step. This last-layer approach was applied in the DeepHF study by fine-tuning the model trained on the large DeepHF dataset on one functional dataset to improve prediction over other functional datasets (Wang et al., 2019). The last-layer approach is commonly used in the field of computer vision, where models are trained on hundreds of thousand of source data, to transfer trained models with many parameters to fit the target datasets. While this TL approach was applied for computer vision tasks, it may be suboptimal for bioinformatics tasks due to the smaller size of the source datasets (Zhang et al., 2018).

In this study, we present a computational method for CRISPR/Cas9 on-target efficiency prediction, called DeepCRISTL, based on TL from our refined version of the source datasets (Ram and DeepHF (Wang et al., 2019). Our newly improved DeepCRISTL model uses a multi-task technique to utilize more enzymes annotated in the DeepHF dataset. In addition, it refines the random ensemble method to improve prediction performance. We trained DeepCRISTL on the high-throughput dataset as the source data, and fine-tuned it on smaller functional or endogenous datasets as the target data. Our evaluations show that DeepCRISTL performs better than the-art methods on all tested datasets. In a follow-up interpretability analysis, we discovered that while some nucleotide positions are commonly important to achieve high efficiencies in various datasets, other positions vary between datasets. The code and software of DeepCRISTL are publicly available via github.com/Orenstein-Lab/DeepCRISTL.

2 Materials and methods

2.1 Data

In the TL technique, there are commonly two datasets: the source dataset and the target dataset. The source dataset, which pre-training is performed, and the target dataset of interest. In this work, source refers to the dataset on which the network is first trained, and target refers to the dataset on which the network is fine-tuned. Table 1 lists the datasets used in this CRISPR high-throughput. All datasets used in this study are publicly available (Haeussler et al., 2016; Leenay et al., 2019; Wang et al., 2018; Kim et al., 2019) and share the same scale of on-target efficiency, i.e. the efficiencies are in a range of 0 to 1.

| Dataset | Genomic region | Cell type | Size |
|---------|----------------|-----------|------|
| High-throughput | CRISPRoN | Human | HEK293 | 23 902 |
| DeepHF WT | Human | HEK293T | 55 604 |
| DeepHF Esp | Human | HEK293T | 58 167 |
| DeepHF HF | Human | HEK293T | 56 888 |
| Functional-U6 promoter | xu2015Train | Human | HL60 | 2076 |
| chari2015Train | Human | HEK293T | 1234 |
| hart2016Train | Human | Hct116 | 4214 |
| hart2016Test | Human | Hct116 | 4239 |
| mall2016Train | Human | Hct116 | 4526 |
| mall2016Test | Human | Hct116 | 3845 |
| x2015TrainKbm7 | Human | KBM7 | 2076 |
| doench2014-Hs | Mouse | TF1 | 881 |
| doench2014-Mm | Mouse | TF1 | 951 |
| 2016_bgr19 | Human | TF1 | 2333 |
| Functional-T7 promoter | exchstruth | Zebrafish | 1-cell embryos | 17 |
| varshney2015 | Zebrafish | 1-cell embryos | 102 |
| gagnon2014 | Zebrafish | 1-cell embryos | 111 |
| shkumatavaPerrigo | Zebralfish | 1-cell embryos | 62 |
| shkumatavaAngelina | Zebralfish | 1-cell embryos | 17 |
| shkumatavaOncorhynchus | Zebralfish | 1-cell embryos | 84 |
| teetool2012morenoMateos2015 | Mouse | Oocytes | 30 |
| Endogenous | Human | T cells | 1636 |

2.1.1 Source datasets

For our study, we used the high-throughput dataset of DeepHF (Wang et al., 2019). The dataset includes on-target editing efficiencies of three enzymes: the wild-type SpCas9 and two types of highly specific SpCas9 variants, eSpCas9 and SpCas9-HF1 (denoted as WT, Esp and HF, respectively). Each gRNA was tested against all 1656 gRNAs. As far as we know, DeepCRISTL is the only TL-based method to predict on-target efficiencies in this dataset. In addition, in the comparison of different TL approaches, we also used a model pre-trained on the CRISPRoN dataset. The CRISPRoN dataset combines two high-throughput datasets: one of 10 592 gRNAs (Xiang et al., 2021) and the other of 13 354 gRNAs (Kim et al., 2019).

2.1.2 Target datasets

1. Functional: For functional CRISPR/Cas9 datasets, we used the datasets curated by Haeussler et al. (2016). This set includes 18 functional datasets.
2. Endogenous: For an endogenous dataset, we used the data produced by Leenay et al. (2019). This dataset contains on-target efficiencies of 1656 gRNAs. As far as we know, DeepCRISTL is the only TL-based method to predict on-target efficiencies in this dataset.
2.2 DeepCRISTL model architecture

The DeepCRISTL-pre-train <WT/Esp/HF> model is based on the architecture of the DeepHF model (Wang et al., 2019; Fig. 2a). The DeepHF model combines an embedding layer for vectorizing the nucleotide one-hot-encoded representation, and a bidirectional long short-term memory (LSTM) layer for identifying sequence patterns in the gRNA sequence. The bidirectional LSTM is a particular subclass of RNN layer, which unlike standard feed-forward neural networks, has feedback connections making it well-suited for processing and making predictions based on sequence-based data (Barash et al., 2020). Since the last two nucleotides of any gRNA + PAM sequence (occupying in total 23 nucleotides) are GG, the input to the model is the first 21 nucleotides with an additional symbol at the beginning of the sequence to inform the model of the sequence start.

We extended DeepCRISTL-pre-train <WT/Esp/HF > model by DeepCRISTL-pre-train-Multi-Task model. This model utilizes the high-throughput datasets of all three enzymes tested in the DeepHF study (Fig. 2b). The model receives an additional input: a one-hot-encoded vector of size 3 to inform the model which of the three enzyme datasets the specific data point came from (WT, Esp, or HF).

Furthermore, to improve prediction performance, we applied a random ensemble initialization technique (Sagi and Roark, 2018). We trained 10 identical models on the same datasets, but with different random weight initialization (Fig. 2c). To predict CRISPR/Cas9 on-target efficiency of a gRNA, we calculate the average prediction over all 10 models. We refer to the final multi-task ensemble model as the DeepCRISTL-pre-train-Multi-Task model.

2.3 Additional input bio-features

To improve prediction performance, the developers of the DeepHF model added bio-features to the input which we also included. They showed that adding those bio-features improved prediction performance compared to using the sequence information alone: the Spearman correlations increased from 0.8555, 0.8491, and 0.8512 to 0.8670, 0.8624, and 0.8607 for WT, Esp and HF datasets, respectively, when the additional bio-features were added to the RNN architecture (Wang et al., 2019). The DeepHF model receives as input 11 bio-features calculated from the gRNA sequence (Fig. 1). The bio-features include three features of the position accessibility of the secondary structure, four features of the melting temperature, three features of the GC-content information, which is known to be strongly associated with the gRNA on-target editing activity, and three sequence-based features. The 11 bio-features are concatenated to the LSTM input, which then passes to the fully connected layer (Fig. 2d). We calculated all bio-features using a script from the DeepHF GitHub repository, which utilizes the ViennaRNA package (Lorenz et al., 2011).

2.4 Training, hyper-parameter search and evaluation

2.4.1 DeepCRISTL-pre-train model training

We randomly split the data into training and test set with sizes of 80% and 15% based on the 10-fold cross-validation. To fairly compare between the single-task and multi-task models, we used the same partition to training and test set in all comparisons. Since some of the gRNA sequences do not have on-target efficiency values for all three enzymes, the test set was constrained to have on-target efficiency values for all enzymes.

2.4.2 DeepCRISTL model fine-tuning

The functional and endogenous datasets are much smaller; hence, the choice of the specific test set can greatly affect the evaluated prediction performance. To obtain a robust evaluation of prediction performance, we repeated the evaluation procedure five times, each
time using a different partition to training and test sets, and we reported the average and standard deviation of obtained Spearman correlations (Supplementary Fig. S1).

Each time, we randomly split the dataset into 80% training and 20% test sets. To determine the optimal number of epochs, we applied 10-fold cross-validation over the training set. In each of the 10 iterations, we found the optimal number of epochs by early stopping. Then, we set the optimal number of epochs over the whole training set as the rounded average over the 10 optimal numbers of epochs. We combined the training and validation to train again on 80% of the data and evaluate prediction performance on the test set. In each of the five iterations, we fine-tuned the 10 randomly initialized pre-trained models, and finally, used all 10 as our TL ensemble model.

2.5 TL approaches

On each endogenous and functional dataset, we tested four types of TL approaches (Fig. 3):

1. Full: Re-training was performed on all model weights.
2. Last layer: Re-training was performed only on the last hidden layer, and the output layer weights.
3. Gradual learning: Initially, re-training was performed only on the last hidden layer weights with the original learning rate. Then, re-training continued on all model weights with a smaller learning rate.
4. No-embedding: Re-training was performed on all model weights except for the embedding layer.

We added two types of trained models for comparison:

1. No-TL: Using only the initial model that was trained on the high-throughput dataset.
2. No pre-train: Training the model only on the endogenous or functional dataset.

The different approaches represent different trade-offs and combinations of fine-tuning the last hidden layer, only re-training the full model. The gradual-learning approach is a unique combination, which leverages both the pre-trained weights in fine-tuning the last hidden layer and the full model re-tuning on a more refined scale. The no-embedding approach is based on the assumption that the embedding layer models general patterns in the gRNA sequence. Hence, there is no need to retain its weights for a modified representation of the nucleotides. Each different approach can represent trade-offs in terms of runtime, where in general, the runtime is proportional to the number of retained parameters (Supplementary Table S2).

2.6 Interpretability

To gain biological insight behind the mechanism of CRISPR/Cas9 on-target editing for each of the datasets, we visualized the sequence preferences learned by our DeepCRISTL models as sequence logos of the input that attained the maximum on-target efficiency as predicted by the model. We gauged nucleotide importance scores of each of the 21 nucleotides of the input sequence, PAM. The scores were generated by the saliency-map method as previously applied (Lanchantin et al., 2017). We then plotted each nucleotide sequence with its height being the importance score.

To calculate the saliency map, we first generated a 22 × 5 matrix to represent the input model, 21 gRNA nucleotides, each represented as a binary vector of size 5 (4 different nucleotides and a symbol for the beginning of the sequence; Supplementary Fig. S2). We initialized this matrix with a value of 0.25 in each of the rows corresponding to nucleotides to represent an initial uniform input. To avoid the effect of the sequence-start encoding, we kept the first column and first row of the matrix as zero, except for the element in their intersection which represents the beginning of the sequence. Since the bio-features are calculated directly from the input sequence, maximizing the model output with respect to the bio-features will lead to the incorrect association between the sequence and its bio-features. To represent a general bio-features input, we calculated the average value of all bio-features in the dataset and multiplied it by a constant to the model. As a result, the constant bio-features have no effect on the derivatives of the model with respect to the sequence. We then computed the derivative of the model with respect to each of the nucleotide inputs. The derivatives modify the input in direction of the gRNA with maximum on-target editing efficiency. The derivative matrix was multiplied by a constant learning rate of 0.1 and then the result was added to the input matrix. For handling the time-series derivative in the LSTM layer, we used the see_rnn python library, which calculates the derivative through time of the output with respect to the input.

3 Results

3.1 Our newly improved DeepHF model

We developed the DeepCRISTL-pre-train model, a modified DeepHF model and training scheme, to improve on-target efficiency prediction. The original DeepHF study reported the Pearson correlation of on-target efficiencies between the enzymes. All enzyme pairwise on-target efficiency Pearson correlations were between 0.6 and 0.8. Thus, to benefit from the combined correlation as shared feature information, we trained a multi-task version of the model on all three enzymes together. In addition, we utilized a random ensemble initialization technique to increase the robustness of predictions. We gauged prediction performance by Spearman correlation of predicted and measured on-target efficiencies on a held-out test set of 15% of the DeepHF dataset, as was previously done in the original DeepHF study (Wang et al., 2019).

Our multi-task model improved prediction performance over all three enzymes’ high-throughput datasets (Fig. 4). The multi-task version achieved a Spearman correlation of 0.878, 0.874 and 0.865 in cross-validation on the DeepHF dataset compared to the single-task version, which achieved 0.873, 0.871 and 0.860 for the WT, Esp and HF enzymes, respectively. This shows the power gained by combining correlated datasets into a single multi-task model.

The addition of the random ensemble initialization technique using 10 differently initialized models improved prediction performance even further (Fig. 4). The ensemble of random initialized multi-task models achieved a Spearman correlation of 0.887, 0.884 and 0.875 compared to the single-task model, which achieved 0.878, 0.874 and 0.865 for WT, Esp and HF enzymes, respectively. When testing other numbers of randomly initialized models, we...
observed no improvement over 10 models at the cost of an increase in training time (Supplementary Fig. S3). Thus, we chose 10 as the number of randomly initialized models in the ensemble.

3.2 Evaluation of different TL approaches
Once we established that the DeepCRISTL-pre-train model was outperforming the original DeepHF model, we turned to fine-tuning it to predict the on-target efficiency of functional and endogenous datasets. We compared four types of TL approaches: last layer, no-embedding, full and gradual learning, and two baseline models: no-TL and no-pre-train, to choose the best TL approach for the task of on-target efficiency prediction. We tested the different approaches in cross-validation on various datasets, each evaluating on five random held-out test sets of 20% of the data. We gauged prediction performance by Spearman correlation of measured and predicted on-target efficiencies. To avoid any potential data leakage between the training and test sets, we removed from all test sets the gRNAs that were present in either the DeepHF or the CRISPROn datasets, except for the three doench datasets where we removed only the gRNAs that were present in the DeepHF dataset. All three doench datasets had an overlap of more than 80% with the CRISPROn dataset, and thus the removal of shared gRNAs would remove most of their data.

The comparison of different TL approaches shows that the gradual-learning approach achieves the best prediction performance compared to all other approaches when pre-training on the DeepHF study (Fig. 5). The gradual-learning approach achieved the highest Spearman correlation on all datasets. For example, the gradual-learning approach achieved an average Spearman correlation of 0.679 on the doench2014-Hs dataset, compared to 0.636, 0.629, 0.633 of full TL, no-embedding TL and last-layer TL, respectively. Results were observed on all other datasets. Moreover, an even greater improvement by the gradual-learning approach compared to the other TL approaches was achieved when pre-training on the CRISPROn dataset and using the CRISPROn deep-learning model (Fig. 5). Interestingly, the DeepHF pre-trained model achieved better prediction performance than the CRISPROn pre-trained model on all datasets. Thus, we chose gradual learning as the TL approach to compose our final DeepCRISTL models based on the DeepHF pre-trained model. Results on the T7 promoter datasets were much worse and less robust for all TL approaches and compared to methods (Fig. 4, Figs. S4 and S5). We speculate that this is due to the small size of the T7 promoter datasets, which hampers training to properly fine-tune the DeepCRISTL pre-train model.

3.3 DeepCRISTL outperforms extant methods in prediction of endogenous and functional on-target efficiencies
To gauge the ability of DeepCRISTL and other methods to predict CRISPR/Cas9 efficiencies on functional datasets, we compared the Spearman correlation achieved by various state-of-the-art methods on all available functional datasets. For each dataset, we randomly held out a test set of 20% of the data to evaluate prediction performance on it. We reported the average over five such test sets for each dataset. We also reported the performance on the endogenous dataset of Leenen et al. but without comparison with extant methods since its predicted scores by all methods mentioned in the Haeussler et al. (2016) study were not available as part of that study.

DeepCRISTL significantly outperformed all other methods in on-target efficiency prediction (Fig. 6). For example, DeepCRISTL achieved an average Spearman correlation of 0.679 on the doench2014-Hs dataset, while the second-best and third-best were DeepHF and CRISPROn, which achieved a Spearman correlation of 0.561 and 0.565, respectively. DeepCRISTL outperformed other methods in all datasets except on xuu2015TrainHh60 and xuu2015TrainKhm7 datasets, where DeepCRISTL’s performance was comparable with Wang score. The good performance of Wang score is explained by the fact that Wang score was trained on these datasets, and thus it got a fair comparison to DeepCRISTL.

3.4 Visualization of gRNA sequence preferences
To gain insights into the mechanism of gRNA on-target sequence preferences, we visualized the sequence preferences learned by DeepCRISTL using sequence logos. We generated nucleotide importance scores using the saliency-map technique for each of the fine-tuned models (each corresponding to a different dataset; Lanchantin et al. 2017). We then plotted each letter in that sequence with its corresponding importance score. We also generated the sequence preference of the pre-trained model to compare the results before and after TL.

Figure 7 shows the on-target preferences learned by the model in each dataset. The G in position 20 is favored by almost all models, which is consistent with previous findings (Wang et al., 2019). We also observe that T in position 14 is favored in the pre-trained model as well as in six of the fine-tuned models and C in position 18 is favored in the pre-trained model as well as in almost all other models. In general, there is a preference for G’s and A’s overall, and favoring of T’s. Interestingly, there are clear differences in the sequence preferences, leading us to speculate on the importance of the cellular context for on-target editing.

4 Discussion
In this study, we developed a novel method, DeepCRISTL, to predict endogenous and functional on-target efficiencies based on TL from high-throughput datasets. The method combines all three enzymes of the DeepHF study and train a joint multi-task model. In addition, we applied an ensemble of 10 randomly initialized models to form one robust model with improved prediction performance compared to a single model. We then fine-tuned the DeepCRISTL pre-train model on smaller endogenous or functional datasets. The resulting DeepCRISTL model achieved state-of-the-art results on held-out test sets of the same datasets.

A key feature of DeepCRISTL improved performance compared to previous methods for the task of endogenous and functional on-target efficiency prediction is the gradual-learning TL approach, which is used to fine-tune the pre-trained model weights. The gradual-learning approach led to state-of-the-art performance across all functional and endogenous datasets. In contrast to extant methods that applied TL by fine-tuning only the last hidden layer of the model, DeepCRISTL first applies TL to train only the last hidden layer and then performs another phase of TL where all model weights are being fine-tuned. We expect that as new high-throughput datasets will be produced in greater scale, prediction
Performance on new endogenous datasets will further improve using gradual learning and similar TL approaches (Supplementary Fig. S6).

To interpret the trained models, we applied the saliency-map technique to visualize the principles the model has learnt before and after fine-tuning. By inspecting the generated sequences logos, we observed key positions, such as the G in position 20, that are shared among all datasets, while other positions are unique, implying that there may be specific cellular factors involved in the on-target editing process.
There are several aspects that require further research in the future. First, further improvement may be achieved by including or even removing some additional bio-features. A careful inspection of the selected 11 bio-features in the DeepHF study shows that some features are highly correlated and even redundant (Supplementary Fig. S7), such as the GC count feature and GC $> 10$ and GC $< 10$ features. Another example is six bio-features related to RNA secondary structure free energy, where all pairwise absolute Pearson correlations are greater than 0.5. Using an established feature selection technique can perhaps improve the results and to accelerate the running time of the model. Second, combining epigenetic marks, such as DNA methylation and open chromatin, may improve prediction performance. Since these data are not always available, there is still room for secondarily-based methods or using predicted epigenetic markers (Schreiber et al., 2020). A key challenge in this aspect is how to expand the trained or pretrained model, which is based on high-throughput data, to include cellular and genomic contexts, by additional cellular information, while enabling efficient optimization of the new and previously trained model weights. Third, further improvement can be achieved by combining all functional datasets in one dataset and using it to fine-tune the model as an intermediate step before the final fine-tuning stage. This may be highly challenging for different cell types and species may be too distinct to be easily merged. One way of combining datasets achieved by linear scaling of the on-target efficiencies, as was recently done in the CRISPROn study (Xiang et al., 2021).

Last, we plan to make our method DeepCRISTL easy to use for biologists by developing a webserver that will receive as input a gRNA sequence and predict its on-target editing efficiencies in different functional and endogenous contexts.

## 5 Conclusion

We developed a new method, DeepCRISTL, to predict the on-target efficiencies of CRISPR/Cas9 given a gRNA and PAM sequence. DeepCRISTL’s unique approach utilizes high-throughput datasets of various enzymes by a multi-task model and improves prediction performance using the random initialization ensemble technique. Most importantly, DeepCRISTL utilizes the gradual-learning approach to transform Cas9 editing principles from high-throughput datasets to smaller datasets. DeepCRISTL outperforms state-of-the-art in on-target prediction, and its learned preferences are biologically relevant. We hope to see DeepCRISTL used to predict on-target efficiencies of functional and endogenous experiments, and aspire after similar developments for experimental datasets based on high-throughput sequencing in other biological domains.

### Acknowledgements

We gratefully acknowledge the support of the CRISPR-IL consortium for supporting our research and for giving access to the AWS platform used for part of this work.

### Funding

This research was partially supported by the Israel Science Foundation [grant number 358/21].

### Conflict of Interest

The authors declare no competing interests.

### References

Barshai, M. et al. (2020) Identifying regulatory elements via deep learning. *Annu. Rev. Biomed. Data Sci.* 3, 315–338.

Cui, Y. et al. (2018) Review of CRISPR/Cas9 sgRNA design tools. *Interdiscip. Sci.* 10, 455–465.

Haeussler, M. et al. (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* 17, 1–12.

Kim, H. K. et al. (2018) Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nat. Biotechnol.* 36, 239–241.

Kim, H. K. et al. (2019) SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci. Adv.* 5, eaax9249.

---

**Fig. 7.** DeepCRISTL model interpretation. We used the saliency-map technique to generate the sequence logo for each of the fine-tuned models, which represents the on-target sequence preference learned on each dataset.
Lanchantin, J. et al. (2017) Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. In: Pacific Symposium on Biocomputing. World Scientific, Singapore, pp. 254–265.

Leenay, R.T. et al. (2019) Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. Nat. Biotechnol. 37, 1034–1037.

Lorenz, R. et al. (2011) ViennaRNA package 2.0. Algorithms Mol. Biol. 6, 26.

Sagi, O., Rokach L. (2018) Ensemble learning: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8, e1249.

Schreiber, J. et al. (2020) Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. Genome Biol. 21, 1–18.

Tian, C. et al. (2018). A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. Springer, New York, pp. 268–279.

Wang, D. et al. (2019) Optimizing CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat. Commun. 10, 4712.

Wang, J. et al. (2020) An overview and the potential of machine and deep learning-based CRISPR gRNA design tools. RNA Biol. 17, 1438–1447.

Xiang, X. et al. (2021) Enhancing CRISPR-Cas9 gRNA prediction by data integration and deep learning. Nat. Commun. 12, 4753.

Zhou, Y. et al. (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature 509, 487–491.

Zhuang, F. et al. (2021) A comprehensive survey on transfer learning. Proc. IEEE 109, 43–76.