piCRISPR: Physically Informed Features Improve Deep Learning Models for CRISPR/Cas9 Off-Target Cleavage Prediction

Florian Störtz*, Jeffrey Mak and Peter Minary+
Department of Computer Science, University of Oxford, Parks Road, OX1 3QD, Oxford, United Kingdom

Abstract

CRISPR/Cas programmable nuclease systems have become ubiquitous in the field of gene editing. With progressing development, applications in *in vivo* therapeutic gene editing are increasingly within reach, yet limited by possible adverse side effects from unwanted edits. Recent years have thus seen continuous development of off-target prediction algorithms trained on *in vitro* cleavage assay data gained from immortalised cell lines. Here, we implement novel deep learning algorithms and feature encodings for off-target prediction and systematically sample the resulting model space in order to find optimal models and inform future modelling efforts. We lay emphasis on physically informed features, hence terming our approach piCRISPR, which we gain on the large, diverse crisprSQL off-target cleavage dataset. We find that our best-performing model highlights the importance of sequence context and chromatin accessibility for cleavage prediction and outperforms state-of-the-art prediction algorithms in terms of area under precision-recall curve.

Introduction

The clustered regularly interspaced short palindromic repeats (CRISPR) sequence family was first described in *E. coli* in 1987 [1], but it took until 2007 to recognise it as a part of the viral defense system of most archaea and bacteria [2]. Exogenous viral DNA is cleaved off by specialised nuclease enzymes, coded for on genomic regions which are often adjacent to CRISPR and hence named CRISPR-associated (Cas). Cleaved-off regions are subsequently incorporated into the CRISPR sequences, which act as a viral history of the respective cell, stabilised by the palindromic nature of their saved states which results in stable secondary structures [3]. From there they can be transcribed to crRNA and invading copies of them can subsequently be rendered inactive. Researchers have used this ability for programmable genome editing in many eukaryotic species, complementing strategies such as zinc-finger nucleases (ZFNs, [4]) and transcription activator-like effector nucleases (TALENs, [5]).

We concentrate on the effects of the wild-type Cas9 protein gained from *Staphylococcus pyogenes*. The crRNA which is originally responsible for recognition of a 20bp viral sequence forms an active complex with the tracrRNA, called single guide RNA (sgRNA), of about 50bp length [6]. Homology of the crRNA part with a 20bp region in the genome results in annealing of the sgRNA with one strand of this region, which we call 'target strand'. Binding happens when the interaction of the 3bp protospacer-adjacent motif (PAM) on the opposite, non-target strand with the Cas9 protein is favourable [7]. For *S. pyogenes* Cas9, this is the case for an 'NGG' PAM where N stands for an arbitrary base (A, T, C, G).

After binding has taken place, nuclease-active enzymes within Cas9 can cleave the double-stranded DNA 3bp upstream of the PAM. Due to the stochastic, energy-driven nature of both the binding and the cleavage process, we expect a distribution of cuts over the whole genome, including undesired off-target effects which could possibly have catastrophic consequences, such as knocking out tumor suppressor genes like p53 and Rb [8]. With a considerable amount of cleavage prediction algorithms present in literature, we present here a model optimisation framework to systematically probe combinations of model architecture, feature set and encoding that maximise prediction accuracy and generate novel biological insight. We want to provide a one-stop side-by-side benchmark of a representative set of modelling approaches.

Methods

Data Source

In order to achieve maximum transparency and comparability, we use guide-target pairs from the crisprSQL dataset [9] curated by our group. It is a collection of 17 base-pair resolved off-target cleavage studies on Sp-Cas9, comprising 25,632 data points and is larger than most datasets used to train prediction algorithms to date. It contains data on various cell lines, mainly U2OS, HEK293 and K562. We have chosen to not include T-cell data data from [10] in order to avoid introducing a considerable cell line imbalance. Experimental data points containing guide and target loci, sequence, cell line, assay type and cleavage frequency have been completed and enriched by sequence context as well as five epigenetic markers (CTCF: chromatin organisation [11], DNase: chromatin accessibility, RRBS: DNA methylation, H3K4me3: histone methylation [12], DRIP: R-loop formation in transcription [13]). We combine these established features for cleavage prediction with physically informed features pertaining to the RNA-DNA heteroduplex in

* florian.stortz@cs.ox.ac.uk, + peter.minary@cs.ox.ac.uk

order to more faithfully map the CRISPR/Cas9 system. We use empirical binding energy estimates given in crisprSQL based on CRISPRspec [14], where the authors calculate four empirical energy contributions:

- binding energy $\Delta G_H^{\text{RNA:DNA}}$ between gRNA and the (off-)target DNA, weighted by a position-wise estimate of the Cas9 influence in the binding,
- free energy $\Delta G_O^{\text{DNA:DNA}}$ of the DNA duplex in the target region required to open the dsDNA,
- free energy $\Delta G_U^{\text{RNA:RNA}}$ of the gRNA (first 20nt) folding computed using RNAfold,
- correcting factor δ_{PAM} determined by the three-letter PAM sequence of the (off-)target.

These can be used to calculate the total binding free energy as

$$\Delta G_B = \delta_{\text{PAM}}(\Delta G_H^{\text{RNA:DNA}} - \Delta G_U^{\text{RNA:RNA}} - \Delta G_O^{\text{DNA:DNA}}).$$

From these we extract the free energy of the RNA-DNA heteroduplex $E_{\text{RNA-DNA}}$, the free energy gained during the RNA-DNA binding starting from an unwound DNA $E_{\text{RNA-DNA}}^{\text{corr}}$ and the free energy released during folding of the sgRNA E_{gRNAfold} to be included in our model (see Figure 1) as

- $E_{\text{RNA-DNA}} = \delta_{\text{PAM}} \Delta G_H^{\text{RNA:DNA}} = E_4$,
- $E_{\text{RNA-DNA}}^{\text{corr}} = \delta_{\text{PAM}}(\Delta G_H^{\text{RNA:DNA}} - \Delta G_O^{\text{DNA:DNA}}) = E_2$,
- $E_{\text{gRNAfold}} = \Delta G_U^{\text{RNA:RNA}} = E_1 - E_4(E_3/E_2)$,

where $E_{1..4}$ refer to the energy features from crisprSQL.

We have found that nucleosome organisation related features are likely to enhance model performance beyond the said five epigenetic markers, and therefore include here three base-pair resolved nucleosomal channels: W/S Score [15], Strong-Weak Nucleotide BDM (SW-BDM) [16], NuPoP (Occupancy) [17]. These scores showed the largest Spearman correlation[1] with cleavage activity when considering 13 distinct nucleosome organisation-related scores. The importance of nucleosome presence informed features for successful predictions of some of our best-performing models supports our emphasis on physically informed features. For a more detailed version of these features, we refer to [15, 16, 17].

Labels

In order to achieve comparability between different studies, we perform a nonlinear Box–Cox transformation [18] to transform the cleavage rates to approximate a Gaussian with zero mean and variance $\sigma^2 = 2$, similar to the approach in [19] and [20]. As we have suggested in [21], cleavage activity (CA) values below the lowest reported assay accuracy of 10^{-5} were set to -4, and transformed values subsequently clipped to the $[-4, 4]$ range. For classification, we define the negative class as all data points with CA $= -4$.

[1] Nucleotide BDM and GC count were ignored due to similarity to SW-BDM and low feature complexity, respectively

Data Augmentation

In order to increase the size of the training set, we augmented it by those putative off-target sites along the respective genome which had fewer than seven mismatches to each respective guide sequence, ignoring duplicates. It was ensured that the protospacer adjacent motif (PAM) was either the canonical 5'-NGG-3' characteristic of SpCas9 [22], or one of the noncanonical forms 5'-NGA-3' and 5'-NAG-3' observed in [23]. These augmented data points were considered negative samples, i.e. labelled with CA $= -4$. This yielded 310,142 total guide-target pairs, making the complete data set highly imbalanced. For this work, we concentrated on the 251,854 data points which originated from a human genome or synthetic DNA.

Feature Encoding

In order to be able to train a variety of models, we first encode the features as a single vector per data point (Figure 1A). We call this the 'target-guide encoding' (E0) since target and guide are represented separately. Sequence information of both guide and target are concatenated as consecutive one-hot vectors. This vector is further concatenated with a vector of epigenetic, energy and nucleosomal features. Epigenetic scores are normalised to $[0, 1]$ for individual markers. CRISPRspec energy factors were normalised uniformly such that their overall maxima and minima also are in $[0, 1]$. Nucleosomal features were taken from the respective algorithms without additional normalisation. Exploring latent representations of guide or target is not within the scope of this work, given that it further complicates comparison between models.

Based on the energy-driven nature of binding and cleavage, we hypothesise that mismatched interfaces affect binding in a totally different way than matched interfaces. This has so far not been recognised in detail by off-target prediction algorithms. In order to explicitly include the information at which a given target-guide pair contains a mismatched interface, we introduce the 'target-mismatch' encoding (E1, see Supplementary Figure 1). Here, we split the target sequence into matched and mismatched interfaces, encoding the matched interfaces in a one-hot fashion first and then append the one-hot encoding of the mismatched interfaces. This encoding loses information about the precise type, i.e. the base pair identity of the mismatch.

In order to remedy this, we introduce the 'target-mismatch type' encoding (E2, see Figure 1B), which also encodes the type of a respective mismatch. It gives the most detailed insight into the nature of a

given base-pair interface, but also requires twice as many one-hot vector entries as the encodings above and might therefore be prone to overfitting.

This is addressed by the 'target-OR-guide' encoding (E3, see Figure 1C), which combines the one-hot encoded versions of guide and target using an entry-wise OR operation between the vectors, as first seen in [24].

For convolutional models, we reshape the resulting vector into a matrix, in which the 23 base pairs each have separate channels for sequence, epigenetic, nucleosomal and energy information (Figure 1D). Where a model requires separate inputs for guide and target, we perform this operation twice with only the respective sequence contributing. This type of matrix encoding is inspired by Chuai *et al.* [25].

Model Architectures

Literature contains a wealth of model architectures commonly used to predict CRISPR cleavage. Currently, successful model architectures for learning-based cleavage prediction fall in one of three categories: tree-based methods [19], convolutional neural networks (CNN, [25, 28]) and recurrent neural networks (RNN, [29, 27]). Thus far, a rigorous comparison between these is missing in literature, and some have only been applied to on-target activity prediction [27]. We therefore take successful CNN and RNN architectures present in the field and adapt them to the task of off-target prediction using various encodings of the subset of features described above.

Our CNN model is based on the architecture described in [25] (see Figure 1E). We replaced the first dropout layer by a Gaussian noise layer since we found this to improve validation set prediction benchmarks. Having batch normalisation layers only before and after the Siamese layers increased training stability without compromising training performance. The same was true for applying (leaky) ReLU activation functions after the conjoined (Siamese) layers. The former also contained a dimension-preserving max-pooling layer in between the convolution and ReLU layers. Furthermore, we found that in our case, removing the 256-filters convolutional layer before concatenation of the Siamese channels did not considerably affect training performance but reduced training time. We then adjusted the downstream network accordingly, which involved removing the penultimate convolutional layer and halving the number of filters in the remaining convolutional layers. The result of this is shown in Figure 1E.

The RNN model is based on the bidirectional gated recurrent unit (BGRU) on-target prediction model from [27]. In order to make it usable for off-target prediction, we combine guide and target sequence using an element-wise OR operation as seen in [24]. The size of the epigenetics channel was increased to accommodate epigenetics, energy and nucleosomal features in a matrix encoding (see Figure 1F).

Model Training & Evaluation

Given the imbalance of validated/measured and non-validated/augmented data points, we employ a bootstrapping strategy as suggested in [30], where training batches on average contain equal numbers of both classes. For regression (classification), early stopping is based on the mean squared error (binary cross-entropy) loss on half of the test set, where the other half is reserved for evaluation.

The XGBoost tree model is trained for 70 epochs, where a new training batch of 50,000 points is sampled in each epoch. We chose hyperparameters `eta`=0.5, `colsample_bytree`=0.7, `max_depth`=7 and a test set early stopping patience of 3 epochs after epoch 30.

The CNN models are trained in the same way, with hyperparameters of `batchnorm_momentum`=0.01, Gaussian noise with $\mu = 0, \sigma = 0.01$ and an Adam learning rate of $1 \cdot 10^{-3}$.

The RNN models are trained for 100 epochs, where batches of 10,000 points are sampled each epoch out of a class-balanced sample of 50,000 data points. Replicating the transfer learning approach taken in [27], we then freeze the network up to and including the second fully connected layer on both sequence and epigenetics branch and train the remaining layers for a further 100 epochs. Both training runs are subject to early stopping with a patience of 4 epochs. Dropout probability was set to 0.2 and the Adam learning rate was at $1 \cdot 10^{-3}$.

Exploring the model space

In order to systematically compare the training performance of combinations of (regression/classification, model architecture, feature selection, feature encoding), we performed individual training runs for all possible combinations within this landscape. To make this independent of our choice of data augmentation, we only included experimentally validated data points within the held out 20% of the crisprSQL dataset in the testing set. The results of this can be seen in Figure 2. We then took only the six best-performing models in terms of both Spearman r (for regression) and area under receiver operating curve (AU-ROC, for classification) further into other testing scenarios detailed below. The six best-performing models were spread across all three architecture types, with CNNs yielding the best performance benchmarks and thus providing three of the best-performing configurations. RNNs and gradient-boosted tree models contributed two and one configurations, respectively.

Testing Scenario 1: crisprSQL comparison

This testing scenario uses the whole crisprSQL dataset in an 80/20 split between training and testing. As suggested in [30], we use all validated and non-validated (augmented as described above) points within our dataset.

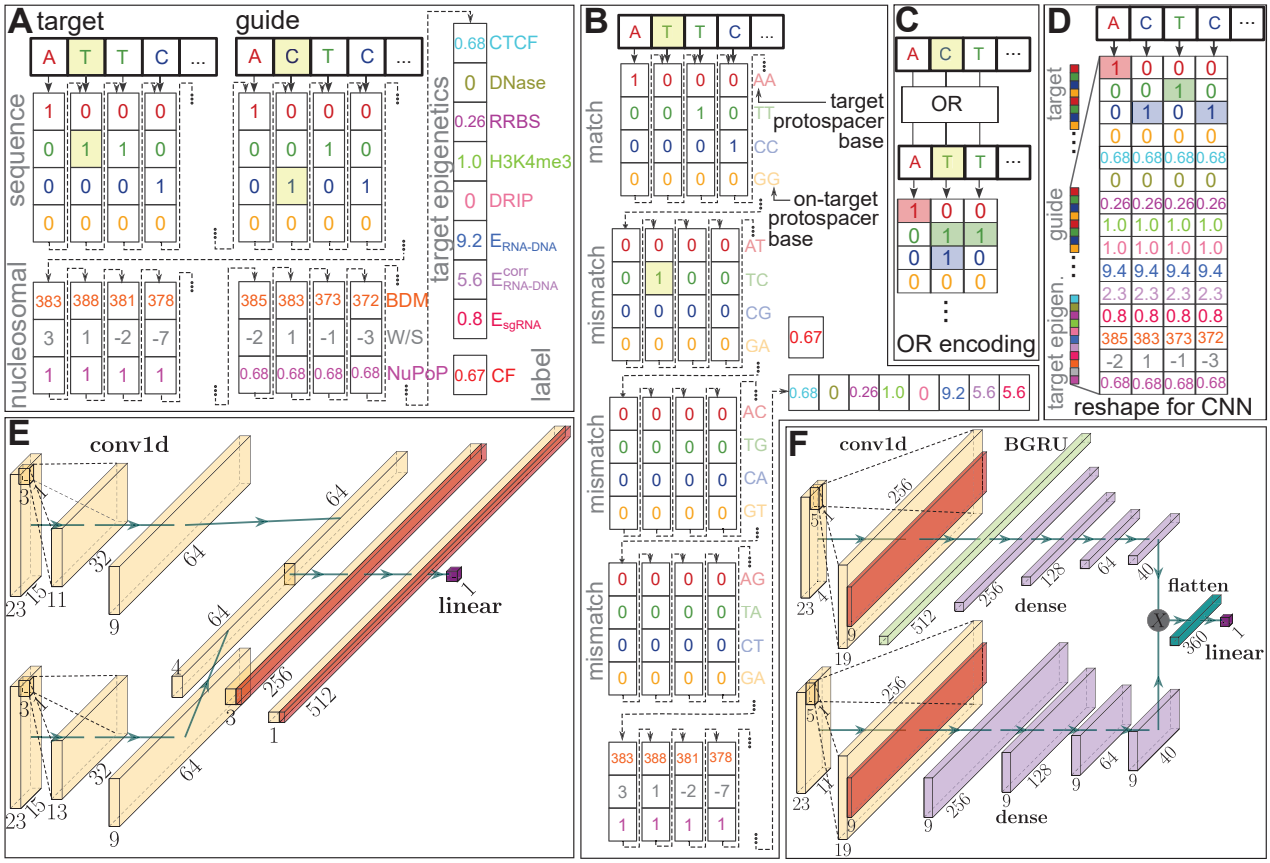
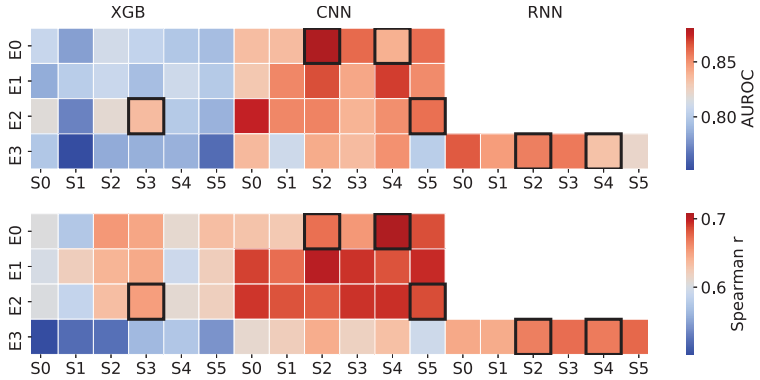


Figure 1: Overview of feature encodings. **A** Target-guide encoding (E0): (Off-)target and guide (on-target protospacer) sequences are one-hot encoded and concatenated together with the epigenetic features (CTCF: chromatin organisation, DNase: chromatin accessibility, RRBS: DNA methylation, H3K4me3: histone methylation, DRIP: R-loop formation during transcription, E: free energy estimates from CRISPRspec [26, 14]) and nucleosome positioning channels (W/S Score [15], Strong-Weak Nucleotide BDM [16], NuPoP Occupancy [17]) of the target to form a single column vector per data point. The Box–Cox transformed cleavage activity (CA) is used as a label. In target-mismatch encoding (E1, see Supplementary Figure 1), the four-letter alphabet is doubled, with the first part encoding the target and the second part indicating whether the base pair on the target strand is canonically matched to the guide, i.e. identical with the protospacer base pair. **B** Target-mismatch type encoding (E2) extends this encoding using a 16-letter alphabet which also contains information about the precise nature of the mismatch. **C** Target-OR-guide encoding (E3) is an efficient way of representing both target and guide sequence in a single four-vector per nucleotide by performing an OR operation between the one-hot encoded target and guide nucleotide vectors (as used in [24]). **D** In order to serve as input for a convolutional neural network, the vector is reshaped into a $(4m + n) \times 23$ matrix where n is the number of non-sequence channels and $m = 1$ for E0 and E3, $m = 2$ for E1 and $m = 4$ for E2. **E** Siamese convolutional neural network architecture used in our CNN model, based on the model in [25]. For encoding E0, target and guide sequence are encoded as separate matrices as shown in panel C and serve as the inputs (left side). Target epigenetics and nucleosomal features are copied to the corresponding channels on the guide arm. For encodings E1–E3, only one arm of the Siamese portion of the network is used, with its number of input channels set to $4m + n$. The channel dimensions given in this figure are valid for the S5 feature set, see Figure 2. **F** Bidirectional gated recurrent unit architecture as used in our RNN model, based on the model in [27]. Sequence features (upper branch) are encoded in E3 encoding; epigenetics-based features (lower branch) are encoded as shown in panel D.

In order to assess the practicality of a large multi-study dataset, we obtained an estimate of the saturation of the data absorption for certain models with respect to the number of studies in training and test data. We gradually introduced training and testing data in ascending order of training performance, i.e. the study with the highest prediction benchmark when training/testing on its 80/20 split being introduced first. The results of this can be seen in Figure 3.

Testing Scenario 2: comparison with other deep learning models

Given that the DeepCRISPR [25], CnnCrispr [31] and AttnToMismatch [28] models all used the combination of datasets [32, 33, 23, 34, 35, 36, 37] as a benchmark, we achieve comparability to the benchmark values quoted there by testing our model predictions on these same studies, with the same imbalance of 1:250 between validated and non-validated points quoted there. All aforementioned studies are included in the crisprSQL dataset, such that we could subsam-



features	name of feature set					
	S0	S1	S2	S3	S4	S5
sequence	✓	✓	✓	✓	✓	✓
energies		✓		✓	✓	✓
epigenetics			✓	✓		✓
nucleosomal					✓	✓

Figure 2: Validation benchmark values for various combinations of feature sets (S0-5, see right panel), encodings (E0-3, see Figure 1) and models (gradient-boosted tree/XGB, convolutional/CNN, recurrent/RNN). Models were trained on a randomly sampled 80% of the crisprSQL dataset [9] and tested on the validated points within the remaining 20%. Rectangles indicate the combinations which we have chosen to scrutinise further.

ple them to use as test set. We reproduced the imbalance by adjusting the bootstrap sampling weights in the validation set appropriately. The results of this are shown in Figure 5.

Pairwise training

When combining various experimental studies, care must be taken as to their interaction during training. We therefore devised the notion of a pairwise training performance of two studies, which we define as the training performance when training on the larger and testing on the smaller of these studies. We did this for all studies in the crisprSQL dataset and visualised the result in a force-directed graph (Figure 4, Supplementary Figure 6) where the force between nodes is proportional to the third power of the Spearman r obtained by pairwise training and testing.

Model Explanation

Using the model-agnostic Shapley value explainer library SHAP [38], we obtain feature importances for each of the input features. Since piCRISPR wraps the feature encoding inside a given model, we retain full explainability of input features even for non-invertible encodings such as E3. Using interface encodings E1 and E2, we obtain an unprecedented, context-sensitive resolution of sequence-based features.

Results & Discussion

Testing Scenario 1

Retaining the aforementioned class imbalance in our dataset in the test set, as suggested in [30], considerably improves regression performance as measured by Spearman r (see Supplementary Figures 4, 5) compared to class-balanced minibatches. An inverse dependence of Spearman correlation on class imbalance ratio raises the question of its suitability when judging a model's regression performance.

Data portioning Figure 3 shows that when gradually increasing training data study by study, with the highest-performing study introduced first, CNN models appear to generalise fastest, whereas RNN models appear to absorb data more gradually. In the present testing scenario (Supplementary Figure 5), this does however not result in an increased test set performance when training on the whole dataset. We also find that all models absorb data more gradually when measured by AUC-PRC than by AUC-ROC or Spearman r, which could support the notion in [30] that AUC-PRC is the more relevant benchmark parameter for this task.

Pairwise training Figure 4 shows that even though studies using the same cell lines show some bunching in a force-directed 2D graph where distance is proportional to generalisation from one study to the other, bunches heavily overlap between cell lines and experimental conditions. This supports the mixing of training data across studies despite differing experimental parameters. Experimental studies conducted on HEK293 cells appear to generalise best. Three distinct studies appear to generalise comparatively poorly to other studies, one of which was gained on HAP1 cells. Study [39] which has been gained on synthetic DNA appears to generalise to other studies about as well as studies on K562 cells.

Testing Scenario 2

Figure 5 shows that all piCRISPR-implemented models achieve state-of-the-art classification performance when measured by AUC-ROC, with RNN models yielding the best results. As mentioned in [30], the area under precision-recall curve (AU-PRC) is a much more suitable measure than the area under receiver operating curve (AU-ROC) for off-target prediction, since in clinical application, false negatives have far more adverse effects than false positives. When considering AUC-PRC as a measure, our RNN_S4E3 model consistently outperforms literature comparison values from references [25, 28, 31]. The S4 feature set (sequence, energy and nucleosomal features) out-

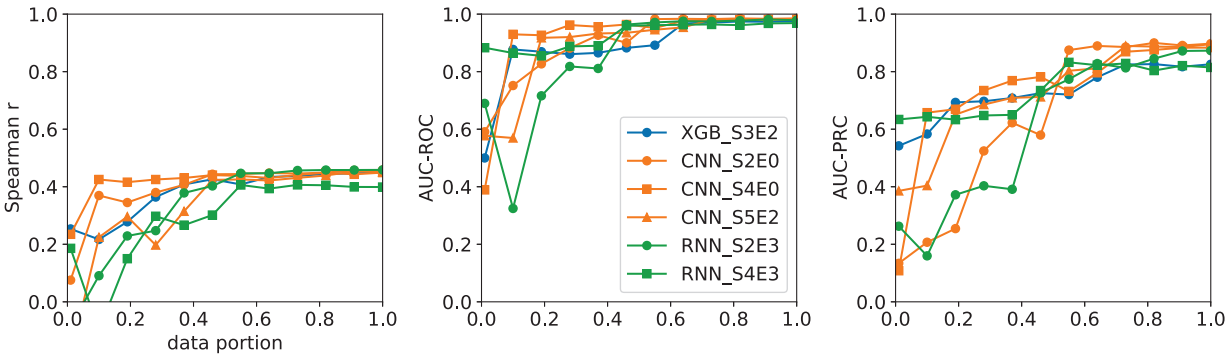


Figure 3: Training performance of the six models when gradually introducing training data study by study along the x-axis, ordered by single-study performance with the highest-performing study introduced first (at low data portion values). Models were trained on a randomly sampled 80% of the crisprSQL dataset [9] and tested on the remaining 20% (testing scenario 1). Points are equally spaced in terms of data portion. The solid line is a guide to the eye. We can see that CNN-based models (orange) seem to generalise fastest, i.e. for lowest data portion values, whereas RNN models appear to absorb data more gradually. Models appear to saturate around the 60% mark, indicating that the weakest-performing 40% of studies might be used for purposes other than training the classifier/regressor.

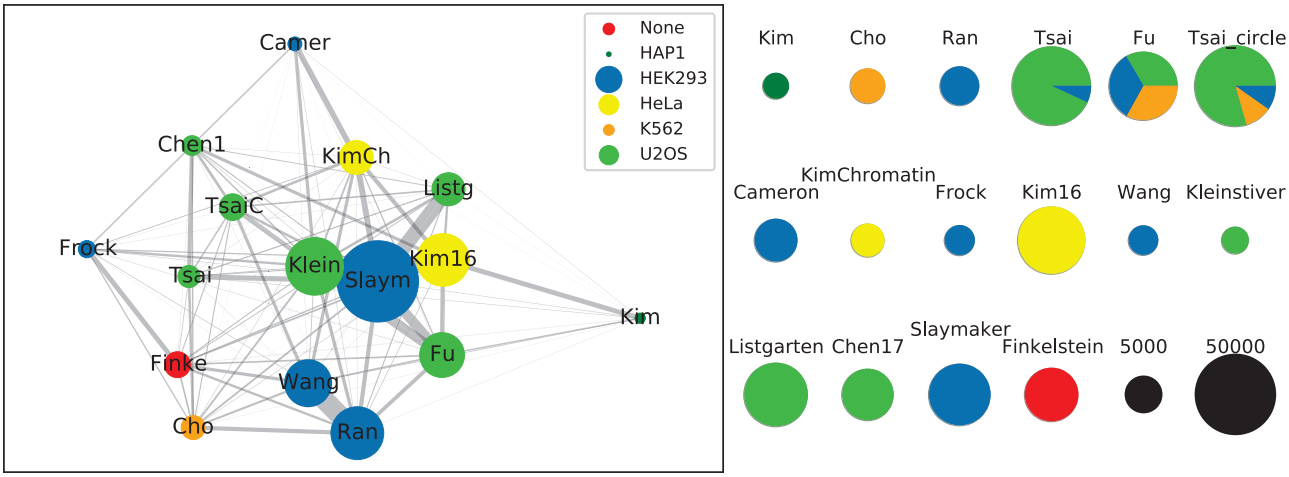


Figure 4: Left panel: 2D distance network representation of pairwise training performance using a Fruchterman-Reingold force-directed algorithm. The force between two nodes (studies) is proportional to the third power of the Spearman correlation when training on the larger and testing on the smaller of these studies. Study labels are abbreviated for better visibility. Edge width represents pairwise training performance of the two adjacent studies alone; bubble size indicates 'study importance', i.e. the overall summed performance of a study. Close positioning indicates good pairwise training performance. Studies have been coloured by majority cell line; all data has been gained using the CNN_S4E0 model which was shown to generalise fastest in terms of Spearman r in Figure 3. **Right panel**: Composition of the cell lines making up each individual study (colours as in left panel) with sizes proportional to the number of data points per study, including non-validated data points. Black circles act as a size legend.

performs the literature-standard S2 feature set (sequence and epigenetics), thereby supporting our hypothesis that nucleosomal features are a key ingredient to cleavage prediction.

When considering regression performance as measure by Spearman r, we find that literature benchmarks from references [25, 31] still outperform the models we have developed in this publication. We hypothesise that this is due to the skewed nature of the cleavage rate distribution and the combination of a large number of studies with differing experimental conditions, which are both only partially remedied by applying the Box-Cox transformation.

Figure 6 shows that the CNN_S5E2 model draws on sequence features which stem from mismatched interfaces differently than on those from matched interfaces, supporting our hypothesis that this differentiation is not only physically indicated but also backed by the model's behaviour. Global SHAP values suggest that the preference of the variable PAM nucleotide at position 21 is contingent on the specific sgRNA–DNA interface formed. Whilst we recover the preference for C and G as well as the disfavouring of T found in literature [40, 25] for matched interfaces, it appears dif-

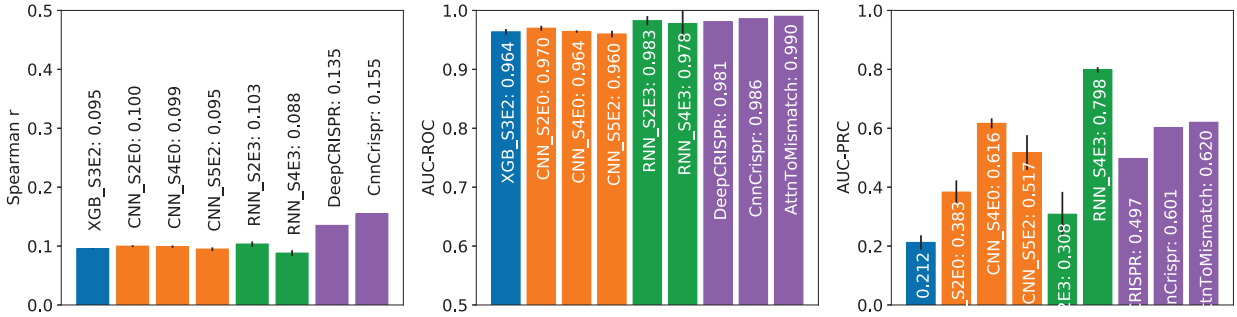


Figure 5: Comparison of models, using various methods, features and encodings, with literature values (purple). The models were trained on a randomly sampled 80% of the crisprSQL dataset [9] and tested on the DeepCRISPR data points within the remaining 20%. Non-validated data points have been oversampled in the test set to match the imbalance in the DeepCRISPR data set (testing scenario 2), such that the benchmark values are comparable with literature results gained on this data set. All models have been implemented in piCRISPR, where training and testing have been repeated 5 times to obtain mean and standard deviation as shown. For the underlying ROC and PRC curves see Supplementary Figure 2.

ferent for mismatched interfaces. The model extracts G to be disfavoured when mismatched with A on the guide, and T to be favoured when mismatched with C on the guide. This supports the notion that a concentration on guide-target interfaces rather than pure base identities is necessary for off-target prediction, and that deeper insight is required than the notion of a preferred base at a specific position. The same goes the notion of a favoured guanine / disfavoured cytosine at position 20 of the sgRNA, which we recover only for specific mismatched interfaces. In accordance with what was found in [25], the model finds a preference of C at position 18 for matched interfaces. We also replicate their result that the RRBS epigenetic feature has the smallest importance of all epigenetic channels. A preference for C and disfavouring of G at position 16 was observed which has been obtained experimentally in [40]. These observations all suggests that it is necessary to consider mismatch interfaces with sequences in the desired genome, not just the sequence of the putative guide, for sgRNA design. We cannot recover the notion of a PAM-distal 'preference zone' and a PAM-proximal 'avoiding zone' of mismatches as seen in [25] when averaging over feature importance values pertaining to each nucleotide, which might be due to our considerably more diverse training dataset. Note that due to the low prevalence of non-NGG PAMs in our dataset, as has been our choice when augmenting it with putative off-targets, the model attributes little importance to the two 5' GG base pairs.

Figure 7 shows that the RNN model, which stands out from literature comparison in terms of AUC-PRC (see Figure 5), is heavily influenced by the SW-BDM feature which is calculated from the 147 base pair region around each nucleotide of the (off-)target DNA. This stands in contrast to the implicit assumption in most cleavage prediction models that sequence context only has a minor effect on cleavage. Our results emphasise the value of the said 147bp sequence context for prediction, since the feature has been shown to carry valuable information relevant to primary chromatin structure [16]. This also demonstrates the importance of nucleosome-related features for cleavage prediction as hypothesised in [21], and supports the notion of chromatin accessibility influencing cleavage activity also found in [41], albeit not across all target nucleotides. To our knowledge, this strong effect of a more than 10bp wide sequence context on genome-wide off-target cleavage prediction has not been demonstrated yet. Hints of it have been seen only for smaller contexts and on-target efficacy prediction [42, 43]. In addition, our findings present an unprecedented example in which the information in the sequence context has the dominant impact on the model.

We further observe that sequence complexity as measured by SW-BDM has a negative influence on the predicted cleavage activity in the PAM-distal region and a positive influence in the PAM-proximal region, with the intermediate region showing a varying but sharply defined importance of the BDM feature. This cannot be explained by the influence of chromatin accessibility on Cas9 binding alone, but hints at further effects of the 147bp sequence context around the target.

As can be seen from the right panel of Figure 7, sequence features play a secondary role in the model. We recover the strong preference for cytosine around base pairs 18-20 observed in [25, 27] as well as the preference for thymine and adenine at position 5 which was observed for on-target interfaces in [25]. Sequence features appear to have little importance in the intermediate region from positions 8 to 14, supporting the notion from [44] that PAM-distal and PAM-proximal mismatches reduce cleavage activity due to cleavage inactivation and abrogation of DNA binding, respectively. We attribute the fact that this model gives little importance to energy features to the nature of recurrent neural networks penalising constant features across the recurrence dimension.

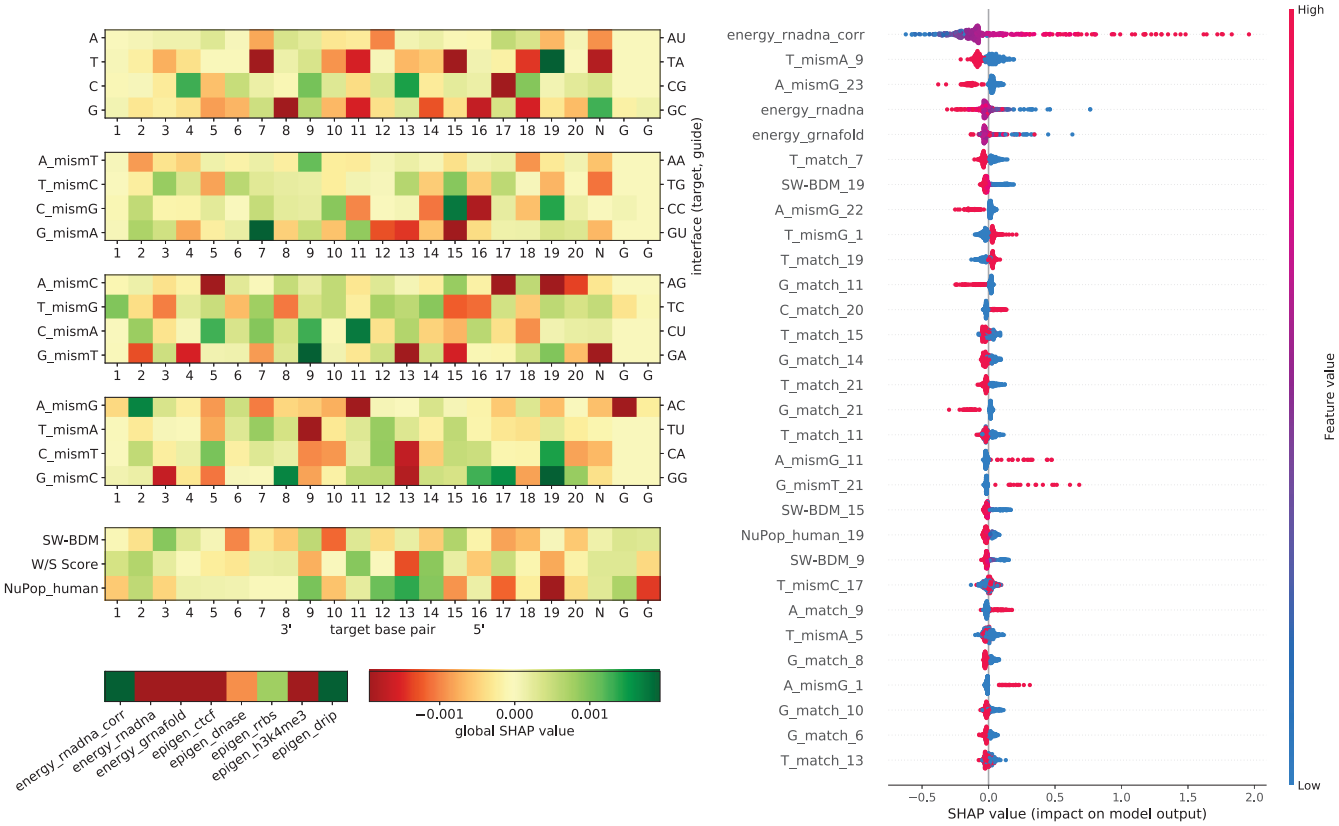


Figure 6: Base-pair resolved global SHAP values for the CNN_S5E2 regression model, where the global SHAP value of a given feature is defined as the mean SHAP value across 500 randomly sampled test set points. Test data consisted of a held out, randomly selected 10% of the augmented crisprSQL dataset. Mismatch channels (middle three heatmaps) are represented by the (off-)target and on-target protospacer nucleotides (left vertical axis) as well as the physical base pair interfaces (right vertical axis), such that `A_mismT` describes all configurations in which an adenine on the target strand faces an adenine on the sgRNA. The bottom heatmap visualises the influence of a set of nucleosomal organisation features on cleavage activity. Negative global SHAP values (red) indicate an average predicted decrease in guide activity for the respective feature. A bar representation of this can be found in Supplementary Figure 7.

Conclusion

Through careful probing of the model architecture, feature selection and encoding space we have identified two neural network-based models which match / outperform state-of-the-art off-target cleavage prediction algorithms. The two models differ in that the convolutional network CNN_S5E2 draws mostly on the sequence nature of the sgRNA-DNA interface, whereas the recurrent network RNN_S4E3 benefits from considering positional features gained from the 147 base pair sequence context around the cut site. While RNN_S4E3 surpasses state-of-the-art models in terms of AUC-PRC, CNN_S5E2 provides insight of unprecedented resolution into the role of a given base-pair interface for cleavage activity.

This reveals several novel, promising pathways towards further improvement of off-target cleavage prediction. Any feature encoding represents a compromise between interpretative resolution and prediction accuracy due to information loss or overfitting from sparsity. We have demonstrated that nucleosome organisation-related features are fit to increase cleavage prediction performance.

Going forward, it could be fruitful to increase model complexity, e.g. using a 2D convolutional (CNN) kernel to capture interaction between features of a single nucleotide. Building on the target-mismatch type encoding (E2), we plan to make use of a 2D convolution kernel which would capture the base-pair resolved interaction between sequence and epigenetic markers as well as between sequence k-mers.

We further envision to replace the epigenetic information of the guide, which so far only copies the epigenetic information of the target DNA. This is clearly an unphysical choice. Given that a synthetic sgRNA does by design not carry epigenetic markers, a one-hot encoded dot-bracket representation of the sgRNA folding gained from ViennaRNA [26] would be a more suitable choice to capture informative properties of the sgRNA.

Data Availability

Models and sample data can be found at github.com/florianst/picrispr.

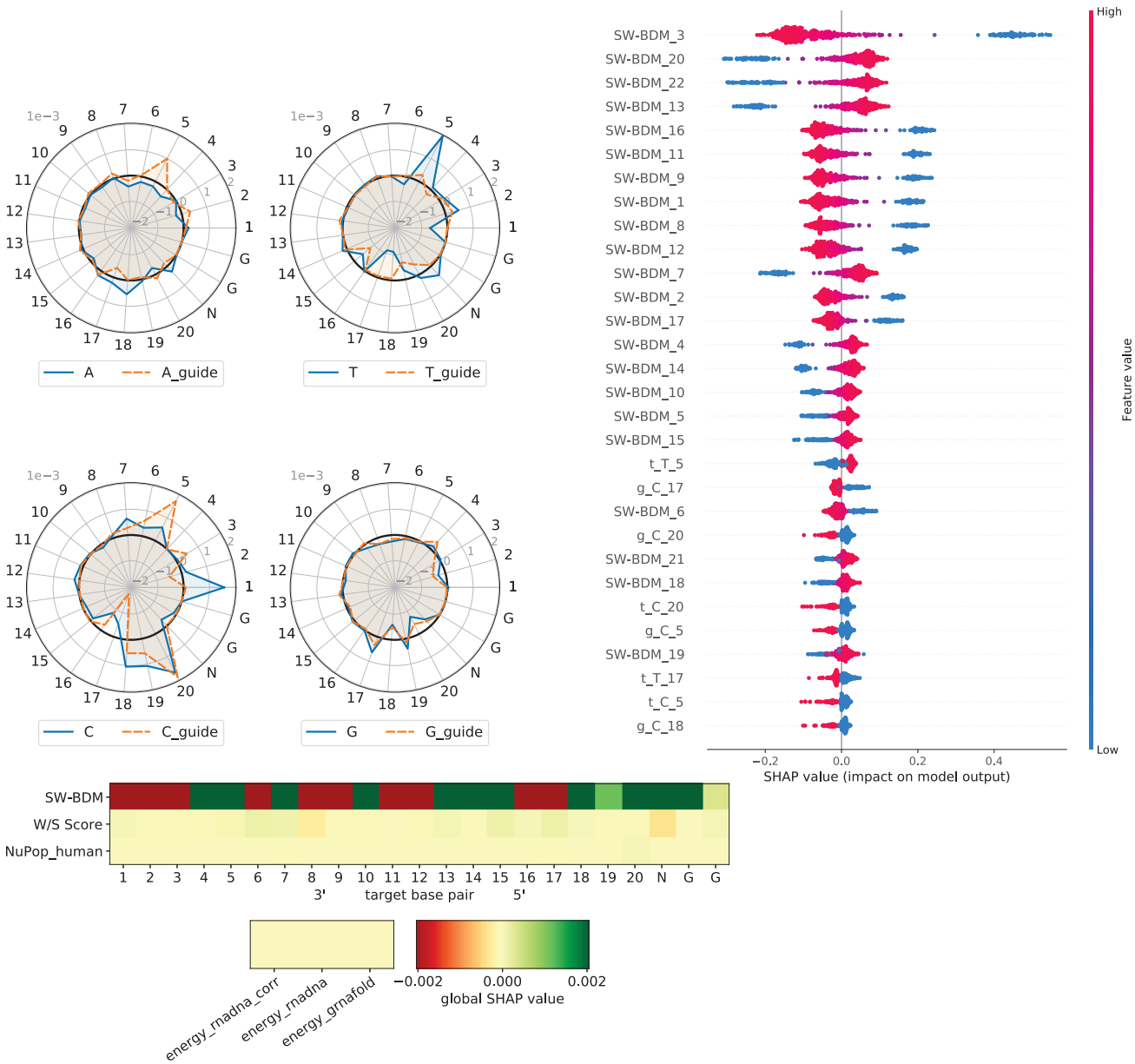


Figure 7: Base-pair resolved global SHAP values for the RNN_S4E3 regression model. Training and test set are identical to Figure 6. Due to the E3 encoding which combines target and guide sequences through a bitwise OR encoding (see Figure 1), target (solid lines in radial plot) and guide (dashed lines in radial plot) feature importances align for most nucleotides. The heatmap shows that the StrongWeakBDM channel contributes the most to the model prediction, with PAM-proximal (distal) nucleotides showing positive (negative) correlation with cleavage activity. PAM-distal and PAM-proximal sequence features inform the model prediction more than intermediate sequence features, matching what was observed in [27]. Energy features show an order of magnitude less influence on model prediction than both sequence and nucleosomal features. A bar chart representation of this can be found in Supplementary Figure 9.

Funding

This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/M011224/1, BB/S507593/1]. Some of the presented results have been obtained using the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558).

The authors declare no conflict of interest.

References

[1] Ishino *et al.* J. Bacteriol. **169**, 5429 (1987).
[2] Horvath *et al.* Science **327**, 167 (2010).
[3] Kunin *et al.* Genome Biol. **8**, R61 (2007).
[4] Urnov *et al.* Nat. Rev. Genet. **11**, 636 (2010).
[5] Joung *et al.* Nat. Rev. Mol. Cell Biol. **14**, 49 (2013).
[6] Ran *et al.* Nat. Protoc. **8**, 2281 (2013).
[7] Wang *et al.* Annu. Rev. Biochem. **85**, 227 (2016).
[8] Ozaki *et al.* Cancers **3**, 994 (2011).
[9] Störtz *et al.* Nucleic Acids Res. **49**, 855 (2020).
[10] Lazzarotto *et al.* Nat. Biotechnol. **38**, 1317 (2020).
[11] Franco *et al.* Biol. Reprod. **91**, 1 (2014).
[12] Sims *et al.* Trends Genet. **19**, 629 (2003).
[13] De Magis *et al.* Proc. Natl. Acad. Sci. U. S. A. **116**, 816 (2019).
[14] Alkan *et al.* Genome Biol. **19**, 1 (2018).
[15] Alharbi *et al.* Genomics, Proteomics Bioinforma. **12**, 249 (2014).
[16] Zenil *et al.* Nucleic Acids Res. **47**, e129 (2019).
[17] Xi *et al.* BMC Bioinformatics **11**, 346 (2010).
[18] Box *et al.* J. R. Stat. Soc. B **26**, 211 (1964).
[19] Listgarten *et al.* Nat. Biomed. Eng. **2**, 38 (2018).
[20] Wang *et al.* bioRxiv **605790** (2019).
[21] Mak *et al.* (2021).
[22] Anders *et al.* Nature **513**, 569 (2014).
[23] Kim *et al.* Nat. Methods **12**, 237 (2015).
[24] Lin *et al.* Bioinformatics **34**, 656 (2018).
[25] Chuai *et al.* Genome Biol. **19**, 1 (2018).
[26] Gruber *et al.* Nucleic Acids Res. **36**, W70 (2008).
[27] Zhang *et al.* Comput. Struct. Biotechnol. J. **18**, 344 (2020).
[28] Liu *et al.* PLoS Comput. Biol. **15**, e1007480 (2019).
[29] Lin *et al.* Adv. Sci. **7**, 1903562 (2020).
[30] Gao *et al.* Brief. Bioinform. **21**, 1448 (2020).
[31] Liu *et al.* BMC Bioinformatics **21**, 1 (2020).
[32] Hsu *et al.* Nat. Biotechnol. **31**, 827 (2013).
[33] Tsai *et al.* **33**, 187 (2015).
[34] Frock *et al.* Nat. Biotechnol. **33**, 179 (2015).
[35] Ran *et al.* Nature **520**, 186 (2015).
[36] Wang *et al.* Nat. Biotechnol. **33**, 175 (2015).
[37] Cho *et al.* Genome Res. **24**, 132 (2014).
[38] Lundberg *et al.* In: *Adv. Neural Inf. Process. Syst. 30.* 2017, 4765.
[39] Jones *et al.* Nat. Biotechnol. (2020).
[40] Doench *et al.* Nat. Biotechnol. **32**, 1262 (2014).
[41] Dhanjal *et al.* Genomics **112**, 3609 (2020).
[42] Xu *et al.* Genome Res. **25**, 1147 (2015).
[43] Boyle *et al.* Sci. Adv. **7**, 5496 (2021).
[44] Jiang *et al.* Annu. Rev. Biophys. **46**, 505 (2017).
[45] Kleinstiver *et al.* Nature **529**, 490 (2016).