piCRISPR: Physically Informed 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 which capture the biological environment of the cleavage site, 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 compares favourably with state-of-the-art prediction performance. We further show that our novel, environmentally sensitive features are crucial to accurate prediction on sequence-identical locus pairs, making them highly relevant for clinical guide design. The source code and trained models can be found ready to use at github.com/florianst/picrispr.

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].

We noticed that repositories of off-target cleavage data contain a significant amount of data points which match in both guide and (off-)target sequence and differ only in the biological environment of the respective loci (see Figure 1). Capturing this environment is therefore instrumental in providing accurate predictions of cleavage activity.

With a considerable amount of cleavage prediction algorithms present in literature [9, 10, 11, 12, 13], we present here a model optimisation framework to systematically probe combinations of model architecture, feature set and encoding. We want to provide a one-stop side-by-side benchmark of a representative set of modelling approaches that take into account physically inspired and environmentally sensitive features. Besides improving prediction accuracy and capturing off-target effects that might so far have gone unnoticed, this will also generate insight into the biological environment that influences CRISPR cleavage.

Methods

Data Source

In order to achieve maximum transparency and comparability, we use guide-target pairs from the crisprSQL dataset [14] 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 [15] in order to avoid introducing a considerable cell line imbalance. Furthermore, the evaluation of our modelling on on-target datasets is beyond the scope of this work due to their different underlying experimental techniques and cleavage quantification measures.

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 [16], DNase: chromatin accessibility, RRBS: DNA methylation, H3K4me3: histone methylation [17], DRIP: R-loop formation in transcription [18]). We combine these established features for cleavage prediction with physically informed features pertaining to the RNA-DNA heteroduplex in order to more faithfully map the CRISPR/Cas9 system. We use empirical binding energy estimates given in crisprSQL based on CRISPRspec [19], where the authors calculate four empirical energy contributions:

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

- 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

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

as the total binding free energy. We extract the binding energy E_{binding} between gRNA and DNA, the free energy E_{DNA} released when opening the dsDNA and annealing with the sgRNA to form the heteroduplex, as well as the free energy released during folding of the sgRNA E_{gRNAfold} to be included in our model (see Figure 2) as

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

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

Besides these established feature, we propose the usage of nucleosome organisation-related features which add an unprecedented level of sensitivity towards the biological environment of the cleavage site (see Figure 1C). We trained a preliminary cleavage prediction model on 13 distinct nucleosome organisation-related scores all based on the 147 bp context around each (off-)target nucleotide (see Supplementary Material) as well as the four literature-standard epigenetic markers named above and include the three scores of highest feature importance: GC count, Nucleotide BDM [20] and NuPoP Affinity [21].

Data Augmentation

In order to increase the size of the training set, we extend it by those putative off-target sites along the respective genome which had fewer than seven mismatches to each respective guide sequence, omitting the (off-)target locus itself. 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]. If a genome-wide off-target detection method has not detected cleavage at a locus within the genome that satisfies these criteria, we deem the cleavage activity at this point to be zero. This yielded 310,142 total guide-target pairs, making the complete data set highly imbalanced. Sticking with the convention in literature, we refer to this process of extending the number of data points as *data augmentation*. For this work, we concentrated on the 251,854 data points originating from a human cell line or synthetic human DNA.

Labels For classification, we define the negative class as all data points with cleavage activity (CA) values below the lowest reported assay accuracy of 10^{-5}, combined with the set of putative off-targets. In order to achieve comparability between different studies for regression tasks, we perform a nonlinear Box–Cox transformation [24] to transform the cleavage rates to approximate a Gaussian with zero mean and variance $\sigma^2 = 2$, similar to the approach in [25] and [11]. Cleavage activity values below the lowest reported assay accuracy of 10^{-5} as well as putative off-targets were set to $-2\sigma^2 = -4$, and transformed values clipped to the $[-4, 4]$ range. This is an empirical choice based on the shape of the resulting distributions.

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 2A). 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 markers (as was done in [9]) which are normalised to $[0, 1]$ for individual markers. We further concatenate this with CRISPRspec energy and nucleosomal features. The former 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 Figure S2). 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 2B), which also encodes the type of a respective mismatch. It gives the most detailed insight into the nature of a given base-pair interface. The 'target-OR-guide' encoding (E3, see Figure 2C) combines the one-hot encoded versions of guide and target using an entry-wise OR operation between the vectors, as first seen in [26].

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 2D). Where a model requires separate inputs for guide and target, we perform this operation twice with only the respective sequence contributing. This type of encoding is inspired by Chuai *et al.* [9].

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 [25], convolutional neural networks (CNN, [9, 10]) and recurrent neural networks (RNN, [11, 28]). Thus far, a rigorous comparison between these is missing in literature, and some have only been applied to on-target activity prediction [28]. 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 comparable to the architecture described in [9]. There, the outputs of two separate, convolutional layer-based encoders for guide and (off-)target are concatenated channel-wise (forming the Siamese part of the network) and serve as input for a convolutional classifier (the conjoined part). We have made various adjust-

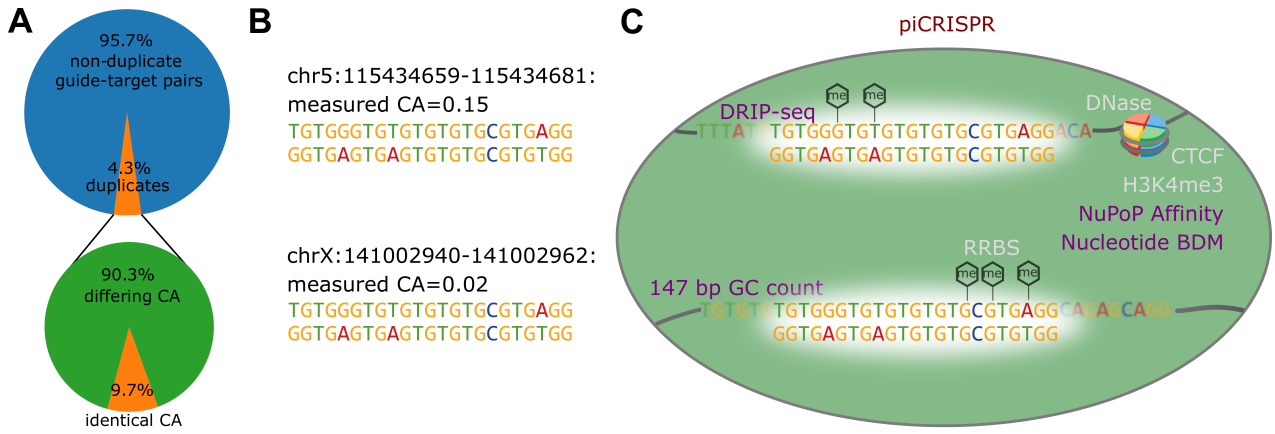


Figure 1: **A** The crisprSQL dataset contains an appreciable amount of perfect guide-target duplicates. We only consider data gained from human cell lines and putative off-targets which we generated based on sequence similarity. 8,922 of 230,274 data points have at least one guide-target duplicate within this set which differs in cleavage activity (CA). As the example from our dataset in panel **B** shows, such a pair looks identical to purely sequence-based prediction algorithms. They might therefore not predict dangerous off-target effects. **C** piCRISPR remedies this by taking into account the biological environment of the cleavage site based on a range of features beyond guide and (off-)target sequence. So far, prediction algorithms have used features related to chromatin organisation (CTCF, [16]), chromatin accessibility (DNase-Seq), DNA methylation (RRBS) and histone methylation (H3K4me3, [17]). We further use features pertaining to R-loop formation during transcription (DRIP, [18]) and the following scores based on the 147 bp sequence context around each (off-)target nucleotide: GC count, sequence complexity (BDM, [20]) and nucleosome positioning information (NuPoP, [21]) which introduce unprecedented sensitivity to the biological environment of the cleavage site. Using these, piCRISPR can correctly rank the two example loci given here.

ments to this architecture based on training stability and validation set performance (see the Supplementary Text). The resulting architecture is shown in Figure 2E.

Our RNN architecture is modelled after the bidirectional gated recurrent unit (BGRU) on-target prediction model from [28]. Here, a BGRU layer is used to make use of the relevant longer-range dependencies between sequence features that would go unnoticed by a CNN of manageable kernel size. In order to make this type of architecture usable for off-target prediction, we combine guide and target sequence using an element-wise OR operation as seen in [26] (see Figure 2C). The size of the model's epigenetics channel was increased to accommodate epigenetics, energy and nucleosomal features in a matrix encoding (see Figure 2F).

Model Training & Evaluation

Given the imbalance of validated/measured and non-validated/augmented data points, we employ a bootstrapping strategy as suggested in [29], 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 Adam learning rate 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. We replicate the transfer learning approach taken in [28] with adjustments to increase training stability and generalisation performance as detailed in the Supplementary Text. Dropout probability was 0.2 and the Adam learning rate was 10^{-3}.

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

Table 1: Abbreviations for feature sets S0-S5.

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 S1. 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 [29], we use all validated and non-validated points within our dataset. Augmenting the measured data points with putative off-targets leads to a class imbalance of 1:10.85 (measured:augmented).

In order to assess the practicality of a large multi-study dataset, we obtained an estimate of the saturation of data absorption for certain models with respect to the set of studies in training and test data. We gradually introduced

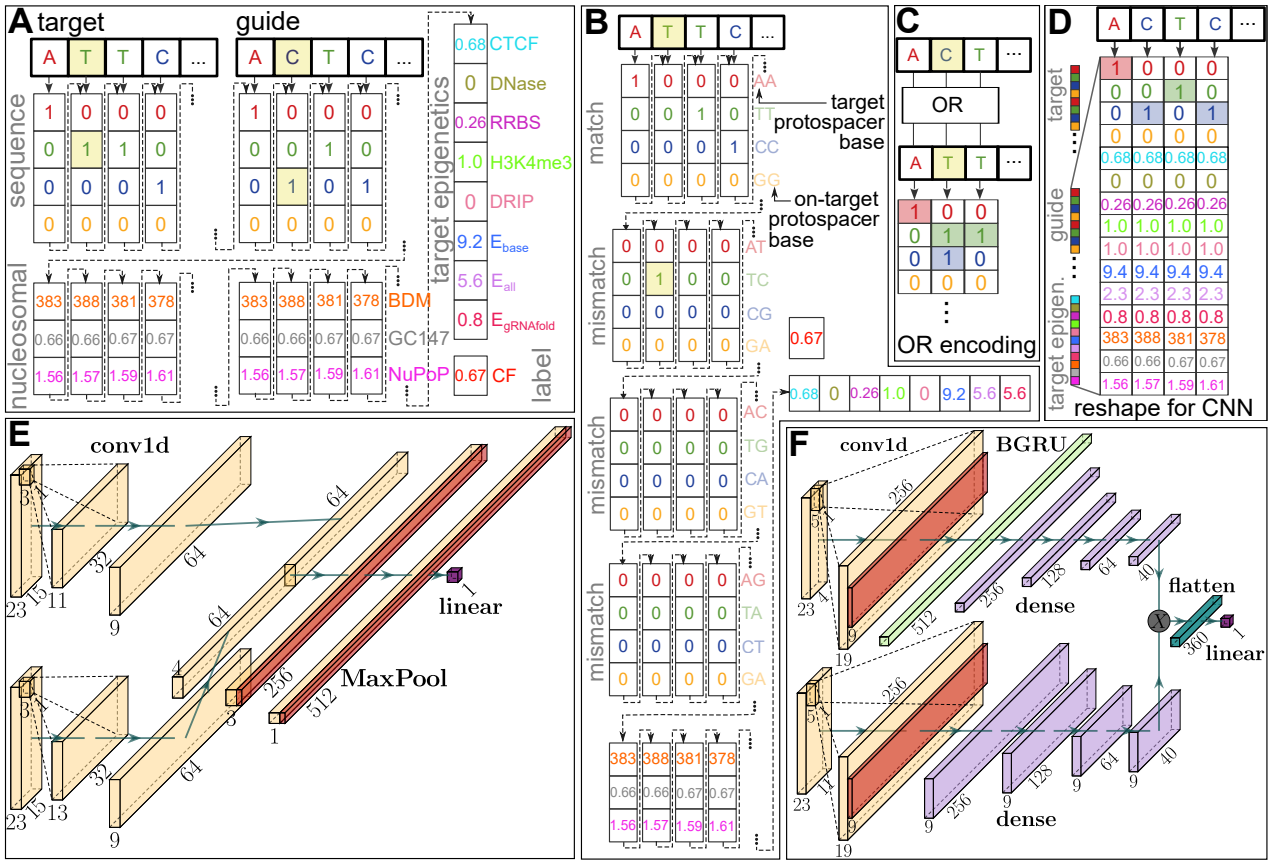


Figure 2: Overview of feature encodings. **A** Target-guide encoding (E0): (Off-)target and guide (on-target protospacer) sequences are one-hot encoded and concatenated together with the epigenetic features (CTCF: chromatin organisation, DNase: chromatin accessibility, RRBS: DNA methylation, H3K4me3: histone methylation, DRIP: R-loop formation during transcription, E: free energy estimates from CRISPRspec [27, 19]) and nucleosome positioning channels (147 bp GC count, Nucleotide BDM [20], NuPoP Affinity [21]) 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 Figure S2), 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 [26]). **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, comparable with the model in [9]. 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 Table 1. **F** Bidirectional gated recurrent unit architecture as used in our RNN model, modelled after the network in [28]. Sequence features (upper branch) are encoded in E3 encoding; epigenetics-based features (lower branch) are encoded as shown in panel D.

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 are shown in Figure S3.

Testing Scenario 2: literature comparison

In this scenario we hold out studies [30, 31, 32] from the training set. These studies have not been included in the training set for the state-of-the-art off-target prediction algorithm CRISPR-Net [11], such that they remain an independent test set to compare CRISPR-Net and piCRISPR side by side. This scenario is our default way of evaluating piCRISPR performance (see Figures 4 and S5). The inherent class imbalance in this test set is 1:103.96.

Testing Scenario 3: set of duplicate pairs

In this scenario, we scrutinise our hypothesis that an environmentally sensitive feature set is fit to not only increase prediction performance overall, but especially for given groups of identical guide-target sequence pairs. To this end we calculate two quantities: First, the mean squared error (MSE) between the predicted regression scores and the ground truth cleavage frequencies within each of the 2,703 groups. Second, the average proportion of the true cleavage activity difference for two points within a given group which the model predicts. This is zero for purely sequence-based models and unity for an ideal predictor. This quantifies how faithful a model is to the differences in biological environment for a given pair. In order to emphasise small deviations which preserve the

rank of predicted cleavage activities, we use the cubic root as a sign-preserving nonlinearity and term this quantity "relative difference". We consider the resulting distributions of both of these quantities for different feature sets.

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 3, Figure S7) where the force between nodes is proportional to the third power of the Spearman r obtained by pairwise training and testing.

Model Explanation

We obtain feature importances using the model-agnostic Shapley value explainer library SHAP [33]. 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.

Sticking with the convention set by the SHAP library [33], we calculate global SHAP values as the mean of the absolute value of SHAP values across data points in the explanation set, which is a random subset of 500 points from the held out test set. In order to show not only the magnitude but the direction in which a given feature influences the model's prediction, we add a sign based on the the average of all non-zero feature values.

Command line usage of our models

We have implemented a command line interface with which piCRISPR predictions can readily be obtained. For maximum usability, the model automatically uses default feature values in case a certain feature was not provided, thereby lowering prediction performance (see Figure S8). The default value of a given feature is defined as the average feature value of the set of those crisprSQL data points which lie within a 20% interval around the mean SHAP value. This means that high-accuracy piCRISPR predictions can be obtained in a user-friendly way, even when providing only guide and (off-)target sequence. Our online repository contains hands-on examples on this.

Results & Discussion

Testing Scenario 1

Data portioning We observe 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. 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 [29] that AUC-PRC is the more relevant benchmark parameter here. Data for this is shown in Figure S3.

Pairwise training Figure 3 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 [34] which has been gained on synthetic DNA appears to generalise to other studies about as well as studies on K562 cells.

	CRISPR-Net [11]			piCRISPR (CNN_S5E2)		
study	[30]	[31]	[32]	[30]	[31]	[32]
AU-ROC	0.991	0.835	0.908	**0.998**	**0.938**	**0.998**
AU-PRC	0.449	0.261	0.246	**0.556**	**0.381**	**0.957**
MCC	**0.335**	0.275	0.280	0.253	**0.424**	**0.493**
F_1	**0.237**	0.305	0.292	0.125	**0.378**	**0.410**
Pearson	0.220	0.276	0.182	**0.317**	**0.340**	**0.318**
Spearman	0.085	0.221	0.179	0.085	**0.258**	**0.214**

Table 2: Comparison between piCRISPR and CRISPR-Net [11]. Both models were tested on the set of studies [30, 31, 32] (testing scenario 2) which neither model has been trained on. We evaluate areas under receiver operating and precision-recall curve, Matthews correlation coefficient (MCC) and the F_1 score as classification benchmarks. For a bar plot representation see Figure S5.

Testing Scenario 2

Figure 4 shows the regression and classification performance of our piCRISPR-implemented models, with CNN_S5E2 yielding the highest benchmarks. As mentioned in [29], 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. An indirect comparison with benchmark values found in literature is shown in Figure S6. When considering AUC-PRC as a measure, the S4 and S5 feature sets, which include nucleosome positioning-related features, outperform the literature-standard sets S2 and S3 which do not (see Table 1). This supports our hypothesis that nucleosomal features are a key ingredient to cleavage prediction.

A direct comparison with prediction results obtained from CRISPR-Net on an identical test set which neither algorithm has been trained on can be found in Figure S5. piCRISPR achieves higher prediction benchmarks for both classification and regression, except for the Matthews Correlation and F_1 score on study [30]. We attribute this to the study's large cell line diversity (see Figure 3). We note that we have observed an inverse dependence of Spearman correlation on class imbalance ratio (data not shown), which necessitates an identical class imbalance ratio in the respective test sets. Our CNN_S5E2 model consistently outperforms CRISPR-Net on the identical held out test set both in terms of Pearson and Spearman correlation.

Feature importance Figure 5 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. We recover the preference for cytosine at position 17 [35, 9] as well as position 20 [9, 28] found in literature for matched interfaces. However, for mismatched interfaces, cytosine is disfavoured. Whilst we cannot recover a strong preference for the variable PAM nucleotide at position 21 for matched interfaces, we observe the preference for guanine reported in literature [35, 9, 28] for

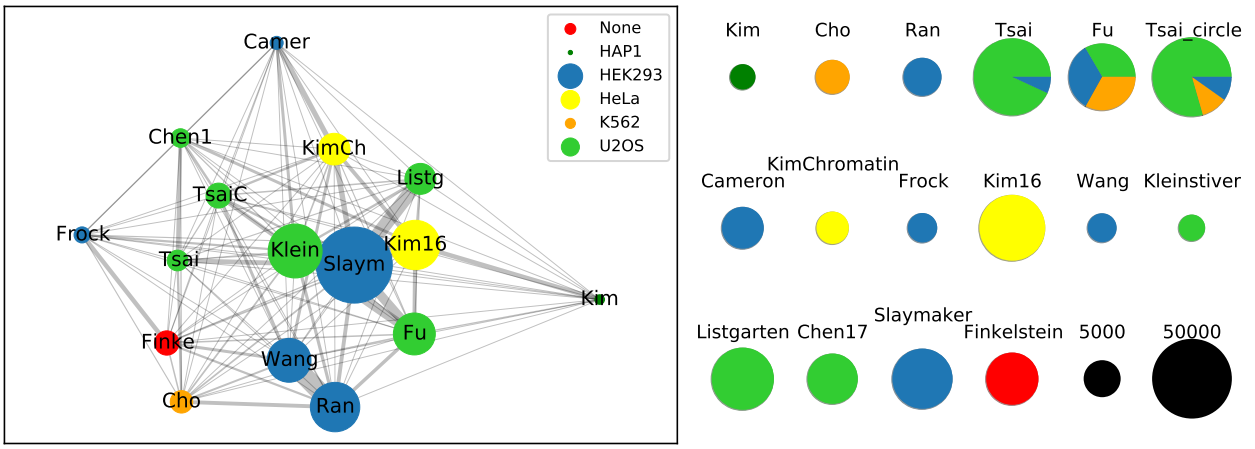


Figure 3: Left panel: 2D distance network representation of pairwise training performance using a Fruchterman-Reingold force-directed algorithm. The force between two nodes (studies) is proportional to the third power of the Spearman correlation when training on the larger and testing on the smaller 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 S3. **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.

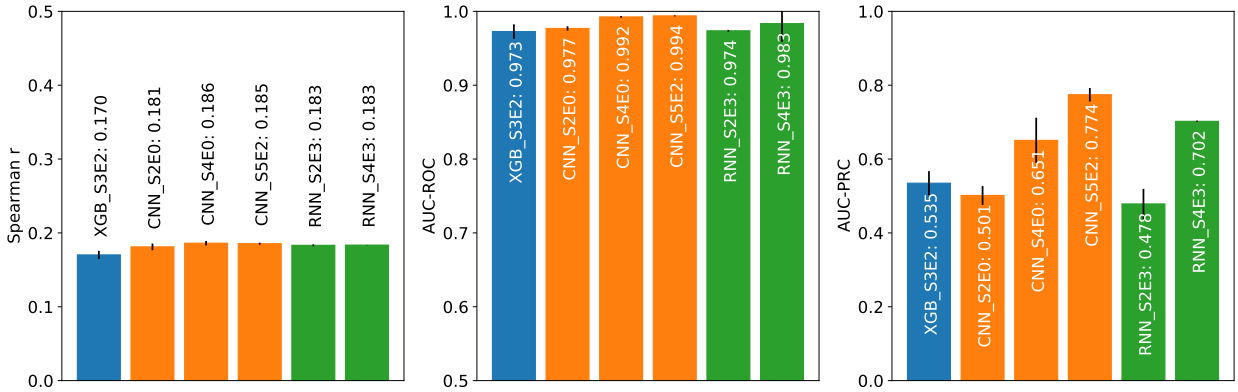


Figure 4: Comparison of models, using various methods, features and encodings. The models were tested on the held out studies [30, 31, 32] and trained on the remaining studies within the crisprSQL dataset [14] (testing scenario 2). Non-validated data points have been oversampled in the test set to match the class imbalance of 1:79.35 found in the dataset I-1 from [11]. 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 Figure S4.

mismatched interfaces. 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. It therefore appears necessary to consider mismatch interfaces together with sequences in the desired genome, not just the sequence of the putative guide, for sgRNA design.

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. We observe the blind spot of mismatch discrimination by the REC3 domain of Cas9 around nucleotide 7 (see also Figure S9) which has been reported in a recent cryo-EM structural study [36] and results in reduced importance of sequence features pertaining mismatched interfaces in this region. At nucleotides 3–5 and 9–11, where mismatch detection by the REC3 domain of Cas9 is high, we observe a mismatch-induced reduction in cleavage activity. We further observe a PAM-distal 'preference zone' and a PAM-proximal 'avoiding zone' of mismatches when averaging over feature importance values by nucleotide, which has been observed in computational [9] as well as cryo-EM [36] studies.

The model draws heavily on two of the empirical energy estimate features, i.e. the folding energy of the sgRNA (E_{gRNAfold}) and the remaining energy contributions during the thermodynamic cycle (E_{DNA}). Considering the largest global SHAP value feature E_{DNA}, we observe a considerable correlation between its value and the SHAP value attributed to it by the model (Figure S11).

When considering nucleosome positioning-related feature channels, we see that the 147 bp GC content around each nucleotide has an overall positive influence on cleavage activity. This supports the prevailing notion in literature insofar as high GC content of the target sequence is on average connected with a GC-rich guide, which have been observed to cause off-target cleavage [35]. We fur-

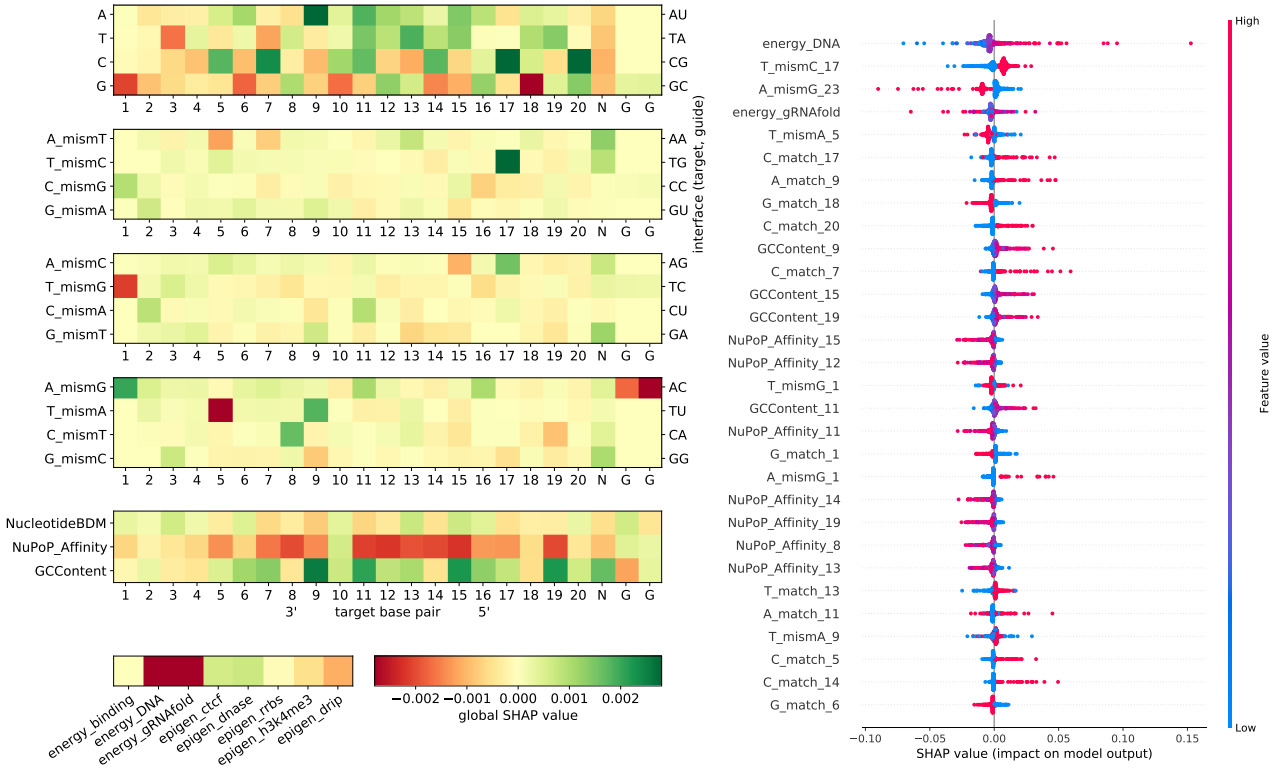


Figure 5: Base-pair resolved global SHAP values for the CNN_S5E2 classification model. Negative global SHAP values (red) indicate an average predicted decrease in guide activity for the respective feature. SHAP values have been obtained on the held out studies [30, 31, 32] from the crisprSQL dataset. Mismatch channels (middle three heatmaps) can be 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 our chosen set of nucleosomal organisation features on cleavage activity. A bar representation of this can be found in Figure S9.

ther observe that the NuPoP Affinity score ranks higher in terms of global SHAP value than most sequence features and all established epigenetic features. The negative influence of nucleosome affinity can be explained by the reduced accessibility of high-affinity DNA regions, and is observed strongly between nucleotides 5 and 19.

This also demonstrates the importance of nucleosome-related features for cleavage prediction, and also supports the notion of chromatin accessibility influencing cleavage activity found in [37]. To our knowledge, this strong effect of a more than 10 bp 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 [38, 39]. In addition, our findings present an unprecedented example in which information in the 147 bp sequence context has considerable impact on the model. A similar analysis for the RNN_S4E3 model can be found in Figures S10 and S12.

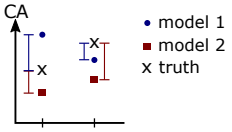
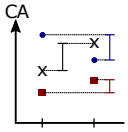
Testing Scenario 3

Table 3 shows that the model performance, measured by the mean squared error of predictions within a group of data points that share both guide and target sequence, is considerably decreased by introducing features beyond sequence information (left column). We find that the lowest error value is achieved when setting the literature-standard epigenetic scores first introduced in [9] to a default value across data points, indicating that they might not interact favourably with other features and not aid our model's absolute predictions. The resulting distribution of MSEs is shown in Table 3. Setting nucleosomal features to a default value leads to an increase in error value, supporting our notion that these contribute to model accuracy.

Looking at the relative pairwise difference, we observe that introducing features beyond sequence leads to an increase of the average proportion of true cleavage frequency differences between points of differing biological environment which is captured by the model. Whilst the full feature set (see Table 1) achieves the highest proportion, setting the epigenetic scores to their default value only has a limited impact on this value, further indicating that their encoding of the biological environment of the cleavage site is of limited importance to the model. In contrast, the proportion drops considerably when setting nucleosomal values to their defaults, supporting the notion that the biological environment of the cut site is captured more accurately by nucleosome positioning-related features (highlighted in Figure 1) than by the literature-standard epigenetic features (greyed out).

Conclusion

Through careful probing of the model architecture, feature selection and encoding space we have identified a neural network-based model (CNN_S5E2) which matches the performance of state-of-the-art off-target cleavage prediction algorithms in direct comparison. It is highly influenced by nucleosome organisation-related features such as histone binding affinity, which emphasises the importance of the biological environment around the cleavage site. Using multiple independent approaches, we have shown that the selection of nucleosome organisation-related features we have newly applied to the task of cleavage prediction informs model predictions more than the epigenetic scores used to date. We have further provided an accessible, user-friendly command line interface that allows users of various disciplines to utilise all our models, even without

		$(\text{pred}_i - \text{truth}_i)^2$ prediction MSE	$\left(\frac{\text{pred}_i - \text{pred}_j}{\text{truth}_i - \text{truth}_j}\right)^{\frac{1}{3}}$ rel. pairwise difference
CRISPR-Net	sequence-only	 0.122	 0.000
piCRISPR	sequence-only	0.139	0.000
	sequence + epigenetic	0.085	0.140
	sequence + nucleosomal	**0.064**	0.331
	full feature set	0.080	**0.367**

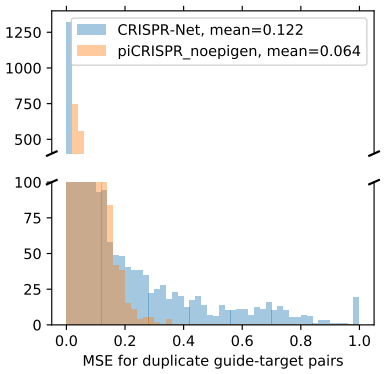


Table 3: Benchmark quantities gained on the subset of duplicate guide-target sequence pairs (testing scenario 3) using our CNN_S5E2 regression model and the CRISPR-Net model [11]. For piCRISPR, we have weakened the model by setting either the literature-standard epigenetic features (DNase, H3K4me3, CTCF, RRBS) or our newly utilised, context-based nucleosome positioning-related features (NuPoP Affinity, Nucleotide BDM, GC count) to a default value across all data points. The non-environmentally sensitive energy and sequence features were retained throughout. **Left column**: mean squared error (MSE) between predicted cleavage score and ground truth cleavage activity, averaged over all groups of identical guide-target sequence pairs. **Right column**: How faithful a model is to the differences in biological environment for a given pair within such a group is measured by the average proportion of the true cleavage activity difference which the model predicts. This is zero for purely sequence-based models and unity for an ideal predictor. To emphasise small deviations which preserve the rank of predicted cleavage activities, we use the cubic root as a sign-preserving nonlinearity and term this quantity *relative difference*. **Right panel**: Example distributions of prediction MSE for the two models. All underlying distributions are shown in Figure S13.

providing a complete set of features. This all paves the way towards the prediction of off-target sites which would so far have gone unnoticed.

Our environmentally sensitive set of features reveals several novel, promising pathways towards further improvement of off-target cleavage prediction. 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 would be a more suitable choice to capture its informative properties.

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] Chuai *et al.* Genome Biol. **19**, 80 (2018).
[10] Liu *et al.* PLoS Comput. Biol. **15**, e1007480 (2019).
[11] Lin *et al.* Adv. Sci. **7**, 1903562 (2020).
[12] Vinodkumar *et al.* Entropy **23**, 608 (2021).
[13] Charlier *et al.* Bioinformatics **37**, 2299 (2021).
[14] Störtz *et al.* Nucleic Acids Res. **49**, 855 (2020).
[15] Lazzarotto *et al.* Nat. Biotechnol. **38**, 1317 (2020).
[16] Franco *et al.* Biol. Reprod. **91** (2014).
[17] Sims *et al.* Trends Genet. **19**, 629 (2003).
[18] De Magis *et al.* Proc. Natl. Acad. Sci. U. S. A. **116**, 816 (2019).
[19] Alkan *et al.* Genome Biol. **19**, 177 (2018).
[20] Zenil *et al.* Nucleic Acids Res. **47**, e129 (2019).
[21] Xi *et al.* BMC Bioinformatics **11**, 346 (2010).
[22] Anders *et al.* Nature **513**, 569 (2014).
[23] Kim *et al.* Nat. Methods **12**, 237 (2015).
[24] Box *et al.* J. R. Stat. Soc. B **26**, 211 (1964).
[25] Listgarten *et al.* Nat. Biomed. Eng. **2**, 38 (2018).
[26] Lin *et al.* Bioinformatics **34**, 656 (2018).
[27] Gruber *et al.* Nucleic Acids Res. **36**, W70 (2008).
[28] Zhang *et al.* Comput. Struct. Biotechnol. J. **18**, 344 (2020).
[29] Gao *et al.* Brief. Bioinform. **21**, 1448 (2020).
[30] Fu *et al.* Nat. Biotechnol. **31**, 822 (2013).
[31] Kim *et al.* Genome Res. **28**, 1894 (2018).
[32] Chen *et al.* Nature **550**, 407 (2017).
[33] Lundberg *et al.* In: *Adv. Neural Inf. Process. Syst. 30*. 2017, 4765.
[34] Jones *et al.* Nat. Biotechnol. (2020).
[35] Doench *et al.* Nat. Biotechnol. **32**, 1262 (2014).
[36] Bravo *et al.* Nature **603**, 343 (2022).
[37] Dhanjal *et al.* Genomics **112**, 3609 (2020).
[38] Xu *et al.* Genome Res. **25**, 1147 (2015).
[39] Boyle *et al.* Sci. Adv. **7**, 5496 (2021).