BMC Genomics

RESEARCH

Open Access

Comprehensive computational analysis of epigenetic descriptors affecting CRISPR-Cas9 off-target activity

Jeffrey K. Mak*, Florian Störtz and Peter Minary*

Abstract

Background: A common issue in CRISPR-Cas9 genome editing is off-target activity, which prevents the widespread use of CRISPR-Cas9 in medical applications. Among other factors, primary chromatin structure and epigenetics may influence off-target activity.

Methods: In this work, we utilize crisprSQL, an off-target database, to analyze the effect of 19 epigenetic descriptors on CRISPR-Cas9 off-target activity. Termed as 19 epigenetic features/scores, they consist of 6 experimental epigenetic and 13 computed nucleosome organization-related features. In terms of novel features, 15 of the epigenetic scores are newly considered. The 15 newly considered scores consist of 13 freshly computed nucleosome occupancy/positioning scores and 2 experimental features (MNase and DRIP). The other 4 existing scores are experimental features (CTCF, DNase I, H3K4me3, RRBS) commonly used in deep learning models for off-target activity prediction. For data curation, MNase was aggregated from existing experimental nucleosome occupancy data. Based on the sequence context information available in crisprSQL, we also computed nucleosome occupancy/positioning scores for off-target sites.

Results: To investigate the relationship between the 19 epigenetic features and off-target activity, we first conducted Spearman and Pearson correlation analysis. Such analysis shows that some computed scores derived from training-based models and training-free algorithms outperform all experimental epigenetic features. Next, we evaluated the contribution of all epigenetic features in two successful machine/deep learning models which predict off-target activity. We found that some computed scores, unlike all 6 experimental features, significantly contribute to the predictions of both models. As a practical research contribution, we make the off-target dataset containing all 19 epigenetic features available to the research community.

Conclusions: Our comprehensive computational analysis helps the CRISPR-Cas9 community better understand the relationship between epigenetic features and CRISPR-Cas9 off-target activity.

Keywords: CRISPR-Cas9, Off-target activity, Nucleosome, crisprSQL, Gene editing, Machine Learning

Background

CRISPR-Cas9 systems are powerful tools for site-directed binding and mutagenesis across a wide variety of eukaryotic species [1–7]. The single guide RNA (sgRNA) in CRISPR-Cas9 is highly programmable and easy to design. As a result, CRISPR-Cas9 has seen its use in many applications. Such applications include

*Correspondence: jeffrey.kelvin.mak@cs.ox.ac.uk; peter.minary@cs.ox.ac.uk

Department of Computer Science, University of Oxford, Parks Road, OX1 3QD Oxford, UK

targeted genome editing, modulation of gene expression [8–10], chromatin visualization [11, 12], epigenetic modifications [13, 14], and chromatin reorganization [15]. Notably, *Streptococcus pyogenes* Cas9 is frequently used due to its short 5'-NGG-3' PAM sequence, which is commonly found in GC-rich mammalian genomes. Nonetheless, CRISPR-Cas9 systems are currently not widely adopted in medical applications, since potential off-target Cas9 endonuclease activity [16–18] may result in undesirable biological effects [19]. To better understand off-target activity, various studies have sought to determine the different factors which influence off-target activity.

A potentially important factor which affects off-target activity is the hierarchical chromatin structure which may block off certain genomic regions. Specifically, previous experimental studies reported less CRISPR-Cas9 cleavage for target sites in heterochromatin compared to those in euchromatin in Cas9 mutagenesis experiments [20, 21]. A similar phenomenon with CRISPR-Cas9 binding activity is observed in dCas9 binding experiments [22–24]. Similarly, chromatin accessibility was observed to positively correlate with CRISPR-Cas9 activity [25, 26]. Chromatin state can be inferred by experimental epigenetic features such as DNase I hypersensitivity, CpG methylation and histone marks. These three features can be experimentally measured by DNase-seq [27], reduced representation bisulfite sequencing (RRBS) [28, 29] and histone ChIP-seq screens [30]. Because of this, various biological studies have used these experimental techniques for investigating the impact of the three epigenetic features (or scores) on off-target activity [31]. In particular, DNase I hypersensitivity and CpG methylation were observed to be highly indicative of dCas9 off-target activity [32]. However, CpG methylation was shown to indirectly contribute to off-target activity. This is because it is the DNA-binding methylation-associated factors which likely block Cas9 binding, rather than CpG methylation [20].

On the computational side, recent deep learning-based CRISPR-Cas9 off-target activity prediction tools [33–35] have used epigenetic features to represent the chromatin state at off-target sites. Such features include CCCTC-binding factor (CTCF, [36]), chromatin immunoprecipitation (ChIP, [37]), histone-3 lysine-4 trimethylation (H3K4me3, [38]), reduced representation bisulfite sequencing (RRBS, [28, 29]) and Deoxyribonuclease-I hypersensitive sites sequencing (DNase-seq, [27]) assays. Available in crisprSQL [39], DNA:RNA ImmunoPrecipitation and high-throughput sequencing (DRIP) is an epigenetic score which measures R-loop formation in the genome [40, 41]. Notably, R-loops play a role in regulating chromatin states [42].

Alternatively, local chromatin structure can be defined as the nucleosome organization at the local region. Nucleosome organization can be described by nucleosome occupancy or nucleosome positioning. Nucleosome occupancy is defined as the cell and time-averaged probability that a given base pair participates in the nucleosomal DNA wrapping around any histone octamer. Nucleosome positioning is defined as the cell and time-averaged probability that a given base pair sits at the center of any 147bp nucleosomal DNA [43]. Nucleosome occupancy is typically measured by Micrococcal Nuclease digestion with deep sequencing (MNase-seq) [44, 45]. Various studies demonstrate that nucleosomes directly inhibit Cas9 binding and cleavage both *in vitro* and *in vivo* [23, 31, 46, 47]. However, access to nucleosomal DNA can be partially recovered via chromatin remodelling [23] and spontaneous nucleosome breathing [47].

In light of the above, we aim to conduct a comprehensive computational investigation on the impact of structural epigenetic features on CRISPR-Cas9 off-target activity. We use the Cas9 off-target activity database crisprSQL [39] and a comprehensive set of computational tools in this study. By doing so, we find that several nucleosome organization-related features attain higher correlation with off-target activity compared to the existing experimental epigenetic scores. In particular, this correlation is significantly higher for two Block Decomposition Method-based features [48, 49]. We also build physically inspired off-target activity prediction models that are purely based on empirical free energy estimates of the sgRNA-DNA heteroduplex and epigenetic features. This allows us to evaluate the impact of epigenetic features in the context of CRISPR-Cas9 activity model prediction. We find that said models take advantage of the computed nucleosome organization-related features but pay less attention to the commonly used experimental epigenetic scores.

Results

Spearman and Pearson correlation analysis

Figure 1 shows two heatmaps denoting the Spearman and Pearson correlations of off-target cleavage activity with 19 epigenetic features (see exact values in Supplementary Table 1). The 19 epigenetic features consist of 6 experimental epigenetic features (names bolded in the figure) and 13 computed nucleosome organization-related features. Heatmap correlations are calculated for target sites in human cell lines HeLa, K562, HEK293 and U2OS from the CRISPR-Cas9 activity cleavage crisprSQL database [39]. To investigate whether correlation values vary between cell lines and genomic regions, heatmap correlations are displayed for all data, individual cell lines and gene/non-gene body regions. The rightmost pie chart shows the cell line composition of the dataset used for analysis.

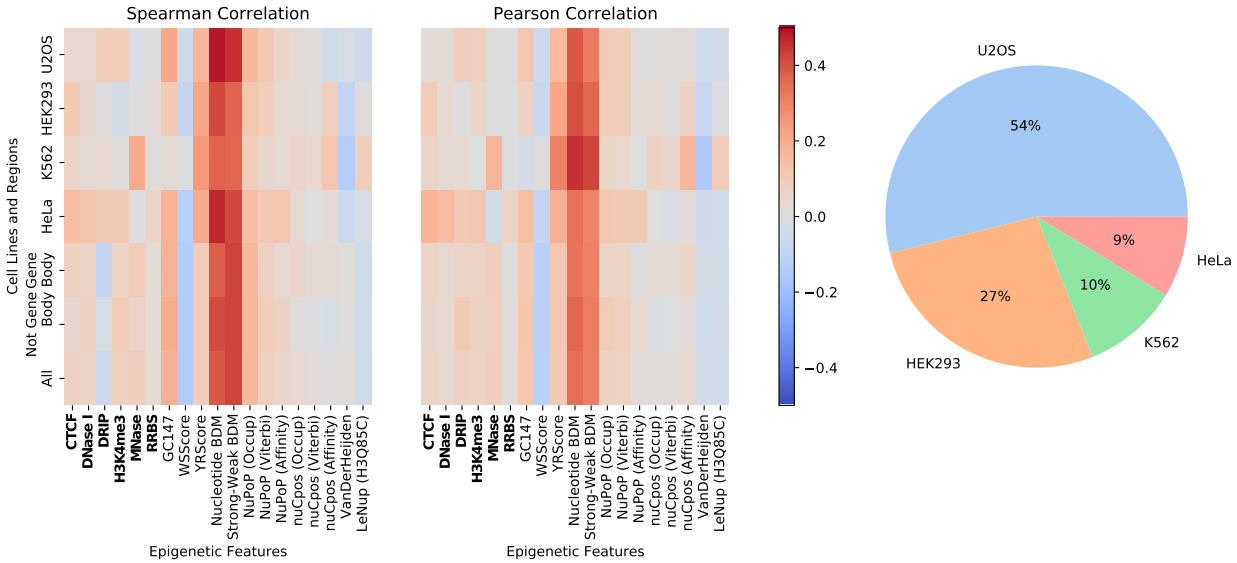

Fig. 1 Heatmaps showing Spearman (left) and Pearson (middle) correlations between 19 epigenetic features and SpCas9 off-target cleavage activities. Red and blue colors represent positive and negative correlations, respectively. The 19 epigenetic features consists of six experimental epigenetic features (bolded) and 13 nucleosome organization-related scores. The first four rows in the heatmaps display cell line-specific correlations. The fifth and sixth row display correlations for off-target sites in gene body and non-gene body regions. The final row displays the overall correlation for the epigenetic features. (Right) Pie plot showing the dataset's cell line composition including all cell lines that contribute more than 1% to the crisprSQL dataset

Table 1 Spearman and Pearson correlation values between SpCas9 off-target cleavage activities and each experimental epigenetic scores for the crisprSQL dataset used in Fig. 1. The experimental epigenetic scores are CTCF, DNase I, DRIP, H3K4me3, MNase and RRBS

Experimental Epigenetic Feature	Spearman	Pearson
CTCF	0.07	0.06
DNase I	0.07	0.03
DRIP	-0.06	0.08
H3K4me3	0.07	0.07
MNase	0.08	0.08
RRBS	0.02	0.01

Overall, Spearman and Pearson correlations for the 19 epigenetic features considered range between -0.5 and 0.5. Only Nucleotide BDM and Strong-Weak BDM, i.e. BDM-based scores, exhibit highly positive correlations when considering all cell lines. Specifically, Nucleotide BDM has Spearman and Pearson correlations of 0.388 and 0.345, and Strong-Weak BDM has correlations of 0.423 and 0.310. Similar values are obtained for Nucleotide BDM and Strong-Weak BDM when considering cell lines individually. When filtering off-target sites by gene body and non-gene body regions, similar Spearman and Pearson correlations are observed across all epigenetic features. This indicates that correlations are not dependent on whether off-targets are in gene bodies. A similar trend is observed when considering each cell line separately (see Supplementary Figs. 2-4).

Table 1 highlights the correlation coefficients for the experimental epigenetic features shown in Fig. 1. In the table, Spearman/Pearson correlations between the six experimental features and off-target cleavage activities in any human cell lines range between -0.1 and 0.1. MNase, which is indicative of nucleosome occupancy rather than nucleosome positioning, has a Spearman and Pearson correlation of 0.08 and 0.08, respectively. Similar values are obtained for the various MNase-seq data across HeLa, K562 and U2OS (see Supplementary Figs. 2, 3 and 4, respectively).

Figure 2 shows the violin and distribution plots for Nucleotide BDM, GC147, YR Scheme and MNase when splitting cleavage activities (CA) into three bins. These bins are CA $= -4$, CA ≤ 2 and CA > 2 (see Supplementary Figs. 5 and 6 for all epigenetic features). In the leftmost column for Nucleotide BDM, most off-target sites with low Nucleotide BDM value fall under the lowest cleavage activity bin CA=−4. The lowest cleavage activity datapoints are almost exclusively composed of augmented datapoints with sequence alignment-derived

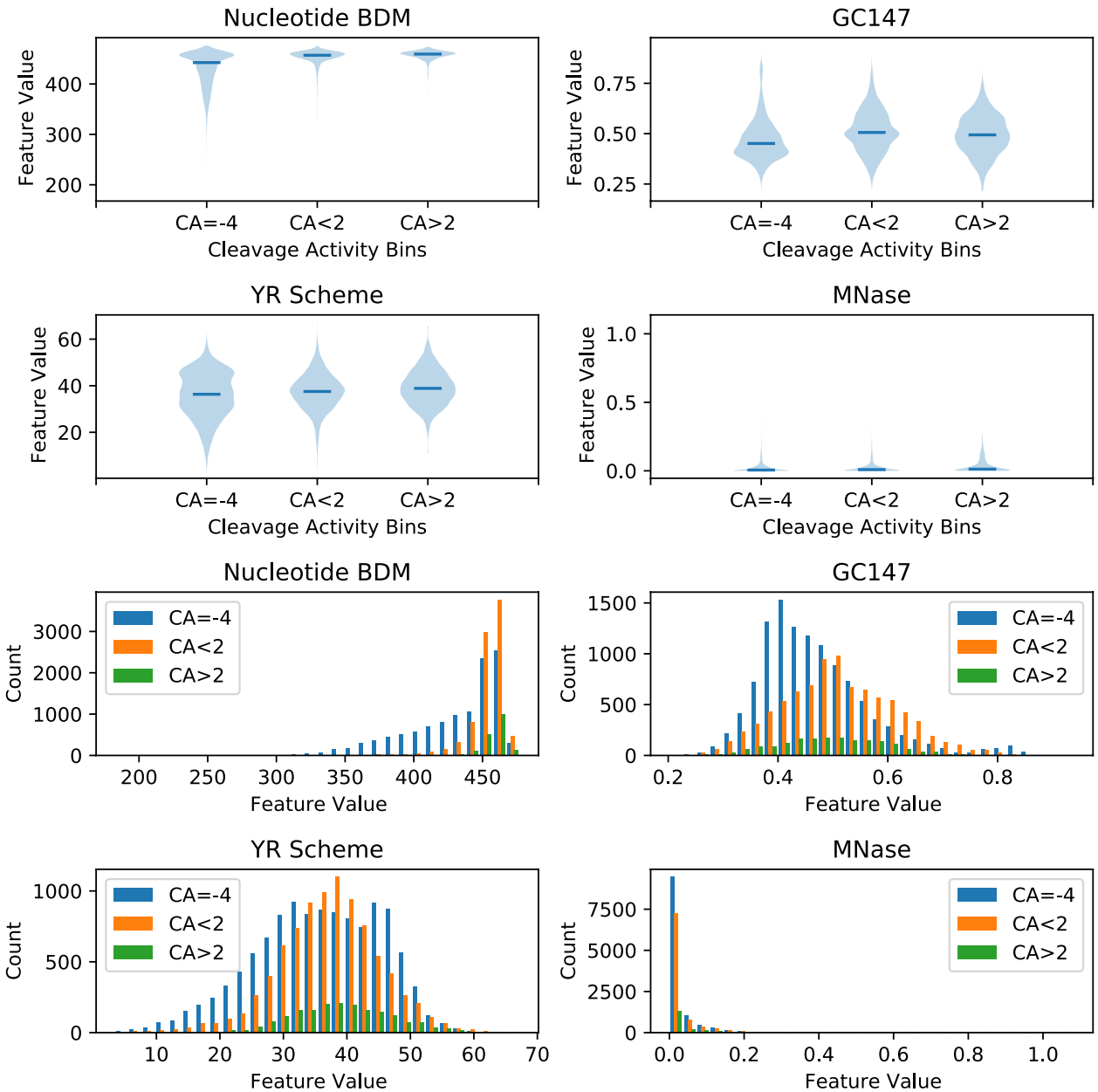

Fig. 2 Violin (top) and distribution (bottom) plots for the epigenetic features with high Pearson correlation, namely Nucleotide BDM, GC147, YR Scheme and MNase. Cleavage activities (CA) are separated into three bins representing low (CA $= -4$, colored blue), medium (CA ≤ 2, colored orange) and high (CA > 2, colored green) cleavage activity. See violin and distribution plots for other epigenetic features in Supplementary Figs. 5 and 6, respectively

putative off-target sites. Such putative off-target sites are assigned the lowest cleavage activity value CA $= -4$ on the assumption that such sites have no off-target activity. Therefore, these datapoints do not carry experimentally derived cleavage activity labels. In addition, these datapoints constitute the larger fraction (52%) of all datapoints. A similar phenomenon is observed for Strong-Weak BDM (see Supplementary Figs. 5 and 6).

Machine/Deep learning-based SHAP analysis

We saw that some computed nucleosome organization-related features correlate with CRISPR-Cas9 off-target activity. As a result, we sought to determine whether the aforementioned features also show patterns in machine and deep learning off-target cleavage activity prediction models. We also sought to investigate the importance of said features without the influence of explicitly encoded

base pair identities. To achieve this, we built two models. The first model is an extreme gradient boosted (XGBoost) tree model. The second model is a convolutional neural network (CNN) model (see Supplementary Fig. 1 for neural network architecture). Both models take all 19 epigenetic features and three binding energy scores as input and predict off-target cleavage activities. Included in crisprSQL, the three energy scores represent free energy estimates used for estimating the DNA-RNA heteroduplex formation's free energy. These energy terms have been generated by using the CRISPRspec [50] biophysical interaction model, which provides various binding energies scores (called CRISPRspec binding energy scores). These binding energies scores are further explained in the Methods section (see CRISPRspec section). The XGBoost and CNN models expect nucleosome organization-related features (scores) at base-pair resolution (23 scores per target site). sgRNA-DNA sequences were not included as input to both models. This is to avoid the interference of sequence features with epigenetic features when computing feature importance scores after training. Instead, we included the sgRNA-DNA sequences-derived CRISPRspec binding energy scores. When testing on the held out 20%, the XGBoost model achieves a Spearman and Pearson correlation of 0.419 and 0.617, respectively. The CNN model yields similar correlations, namely a Spearman and Pearson correlation of 0.424 and 0.594, respectively.

Next, we interpret the two model using SHAP (see Methods section) after training and evaluating the contributions of each input feature. To evaluate a model, a randomly drawn test dataset containing 2000 points is used. Figure 3 and Supplementary Fig. 7 show the resulting feature-based SHAP summary plot and base pair resolution heatmap, respectively, for the trained XGBoost model. An analogous summary plot and heatmap for the CNN model can be found in Fig. 4 and Supplementary Fig. 8. In the two SHAP summary plots, the distribution of SHAP value contributions is shown for every input feature present in the model. Model input features are ordered in decreasing SHAP feature importance. In other words, features at the top carry high SHAP feature importance, and features at the bottom carry low SHAP feature importance. In both SHAP summary plots, the SHAP feature importance of the six experimental epigenetic scores are not comparable to the nucleosome organization-related scores. In addition, the top five scores with highest SHAP feature importance include Nucleotide BDM and NuPoP (Affinity). These two features display similar correlations between feature value and SHAP value across Figs. 3 and 4. Notably, low Nucleotide BDM values and high NuPoP (Affinity) values correspond to negative impact on off-target activity. As for the three CRISPRspec binding energy scores, they attain comparable SHAP feature importance to top-performing nucleosome organization-related scores in both models.

Discussion

MNase-seq, a common genome-wide experimental technique, appear to be an attractive option for obtaining raw nucleosome occupancy data. In addition, nucleosome occupancy data may be indicative of CRISPR-Cas9 off-target activity. On account of this, we sought to obtain MNase-seq data from NucPosDB [58] where available for human cell lines. Nonetheless, we found MNase-seq data only for U2OS, K562 and HeLa in NucPosDB. In particular, MNase-seq data must be measured for each cell line of interest in order to curate sufficient data for analysis. This makes MNase-seq data cell-based and difficult to obtain. Such qualities are the opposite of computed nucleosome organization-related scores, which are not only genome-wide but also easy to obtain and cell-line independent.

The experimental features CTCF, DNase I, RRBS and H3K4me3 are commonly used as input features in multiple state-of-the-art deep learning-based CRISPR-Cas9 off-target activity prediction tools [33, 35, 59]. Despite this, we can see in Fig. 1 and Table 1 that BDM-based scores attain much higher Spearman and Pearson correlations with off-target cleavage activities. This is in contrast to the six experimental epigenetic features listed in Table 1, which do not strongly correlate with cleavage activity.

Scrutinizing the distributions in Fig. 2, we observe that most off-target sites with low Nucleotide BDM value fall under the lowest cleavage activity bin CA=−4. Moreover, crisprSQL is augmented with sequence alignment-derived putative off-target sites. Such putative sites are assigned the lowest cleavage activity value CA= −4 on the assumption that such sites have no off-target activity. As a result, among the putative sites, Nucleotide BDM is better at separating sites without activity from sites with activity, compared to other epigenetic features. The aforementioned observations can be explained by the correspondence between low Nucleotide BDM values and proximity to nucleosome dyad positions [49]. Since these positions are blocked by nucleosomes, they are inaccessible for Cas9 binding and cleavage, thus resulting in low off-target activity. This is a possible explanation on why these off-target sites have not been experimentally identified as active. In practice, the application of Nucleotide BDM for data filtering can be useful when preparing data for CRISPR-Cas9 off-target model training. This is because such a filtering might help resolve any class imbalances between experimentally measured and putative off-targets. Moreover, Nucleotide BDM is a fundamental property of the 147bp nucleosomal

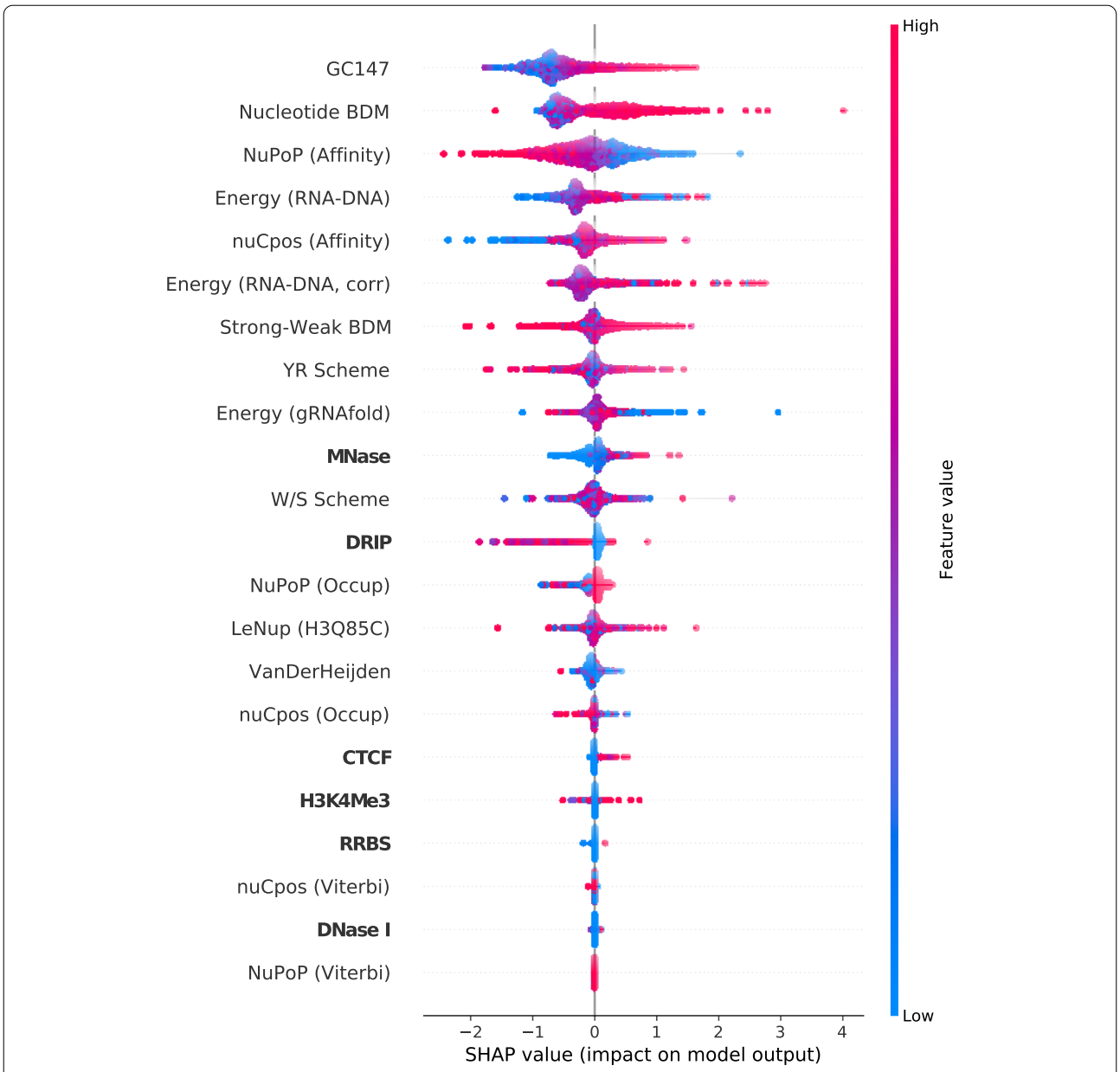

Fig. 3 SHAP summary plot for the trained extreme gradient boosted (XGBoost) tree model. The model's input consists of three CRISPRspec-derived energy terms, six experimental epigenetic scores (bolded), and 13 computed nucleosome organization-related scores. The three CRISPRspec-derived energy terms are $E_{\text{RNA-DNA}}$, $E^{\text{corr}}_{\text{RNA-DNA}}$ and E_{gRNAfold}. The six experimental epigenetic scores are CTCF, DNase I, DRIP, H3K4me3, MNase and RRBS. The 13 computed nucleosome organization-related scores are GC147 [51], W/S scheme, YR scheme [52, 53], Strong-Weak BDM, Nucleotide BDM [48, 49], NuPoP (Occupancy), NuPoP (Affinity), NuPoP (Viterbi) [54], nuCpos (Occupancy), nuCpos (Affinity), nuCpos (Viterbi) [55], VanDerHeijden [56] and LeNup (H3Q85C) [57]. The base pair-resolved SHAP contributions for each data point are summed for each computed nucleosome organization-related score

DNA context which is not dependent on any training dataset. Deeper understanding of why augmented datapoints (i.e., lowest cleavage activity datapoints) have no off-target activity is currently lacking. To the best of our knowledge, there has not been any existing target sequence-based measure that could separate augmented datapoints from experimentally-derived datapoints. Figure 2 indicates that low values of Nucleotide BDM can separate these augmented datapoints remarkably well compared to other similar measures.

In Figs. 3 and 4, the six experimental scores' low SHAP feature importance demonstrates that they are

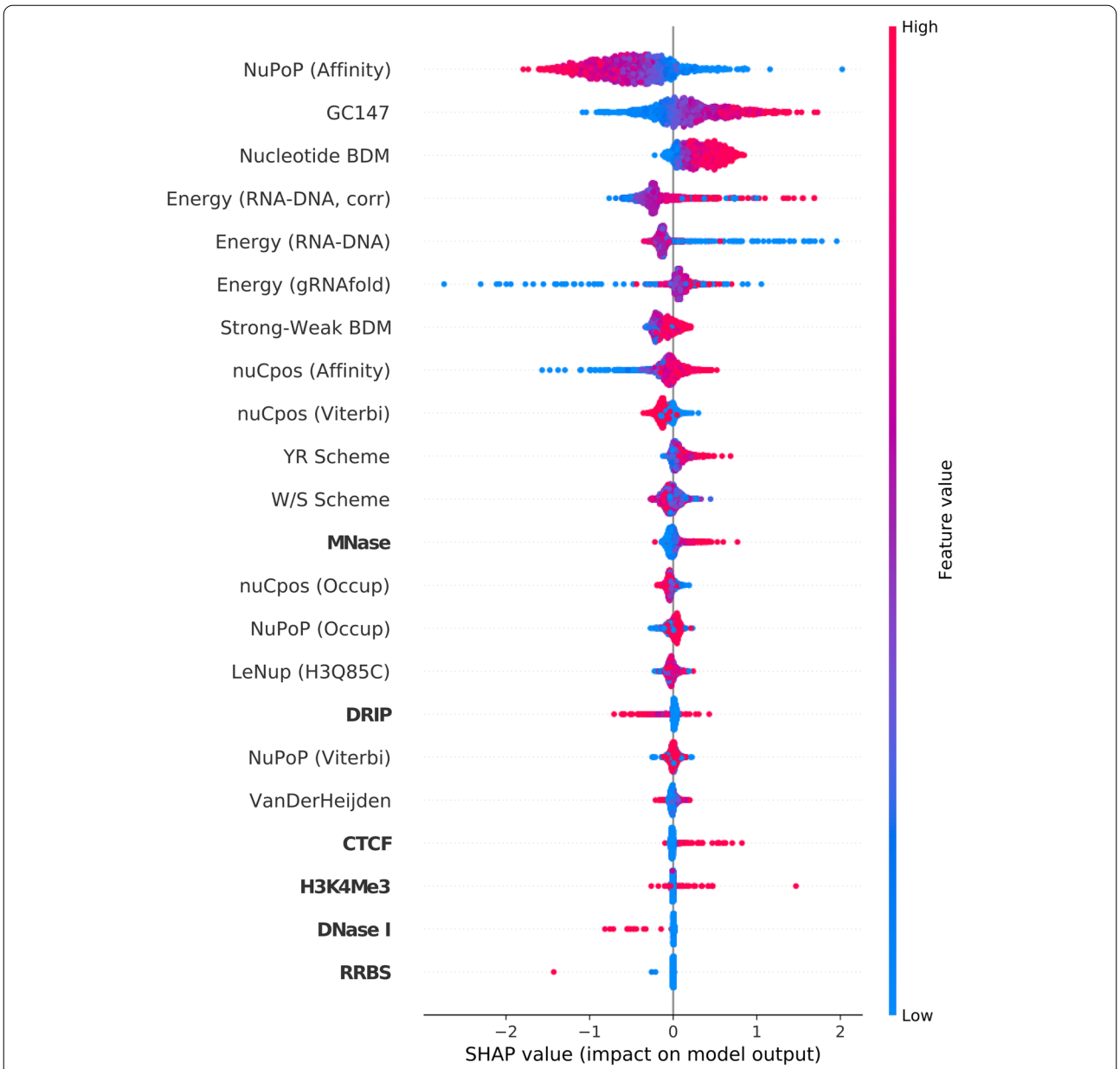

Fig. 4 SHAP summary plot for the trained convolutional neural network (CNN) model (see Supplementary Fig. 1 for architecture details). The model's input consists of three CRISPRspec-derived energy terms, six experimental epigenetic scores (bolded), and 13 computed nucleosome organization-related scores. The three CRISPRspec-derived energy terms are $E_{\text{RNA-DNA}}$, $E^{\text{corr}}_{\text{RNA-DNA}}$ and E_{gRNAfold}. The six experimental epigenetic scores are CTCF, DNase I, DRIP, H3K4me3, MNase and RRBS. The 13 computed nucleosome organization-related scores are GC147 [51], W/S scheme, YR scheme [52, 53], Strong-Weak BDM, Nucleotide BDM [48, 49], NuPoP (Occupancy), NuPoP (Affinity), NuPoP (Viterbi) [54], nuCpos (Occupancy), nuCpos (Affinity), nuCpos (Viterbi) [55], VanDerHeijden [56] and LeNup (H3Q85C) [57]. The base pair-resolved SHAP contributions for each data point are summed for each computed nucleosome organization-related score

inappropriate for informing off-target cleavage activity prediction models. This corroborates with the Spearman and Pearson correlation values in Table 1. The top five scores with highest SHAP feature importance include Nucleotide BDM and NuPoP (Affinity). The two features show similar correlations between feature value and SHAP value across the two plots. Notably, low Nucleotide BDM values and high NuPoP (Affinity) values correspond to negative impact on off-target activity. This observation corroborates the fact that such feature values often are signals of positioned nucleosomes. It follows that information in BDM-based scores and NuPoP (Affinity), alongside other nucleosome organization-related scores, are well suited for informing off-target

cleavage activity prediction models. GC147's importance as a feature in both machine learning models is in agreement with latest findings [60] that CRISPR-Cas9 bends DNA to read its sequence. Specifically, DNA bendability is very highly correlated with GC content [61]. Such a fact could explain the findings of the SHAP summary plot, namely that high GC147 has a positive impact on (off-)target cleavage activity. The three CRISPRspec binding energy scores contribute significantly towards model predictions in both models, which confirms these scores' usefulness for CRISPR-Cas9 off-target activity prediction. Despite interesting structures in the heatmaps of Supplementary Figs. 7 and 8, a thorough analysis of such structures is beyond the scope of this study. In off-target prediction, the most suitable use case for Nucleotide BDM and other relevant measures is to incorporate them in 'complete' deep learning models. Together with measures like NuPoP (Affinity) and GC147, they can be combined with the guide-RNA-(off-)target DNA sequence pair as input to such models.

Interestingly, only BDM-based scores have noticeable correlation with (off-)target activities. However, the NuPoP (Affinity) score has comparable SHAP feature importance to Nucleotide BDM score in both machine learning models considered in this work. Supplementary Fig. 9 shows that the correlation between NuPoP (Affinity) and Nucleotide BDM is relatively low. This observation agrees with the finding that only one of the two scores (Nucleotide BDM) correlates with (off-)target activity. However, it does not alone explain why the other score, NuPoP (Affinity), is still a comparably impactful feature in both machine learning models. To investigate this further, we obtained SHAP dependence plots for both models, which include NuPoP (Affinity) and Nucleotide BDM (see Supplementary Figs. 11 and 12). These plots show that a given NuPoP (Affinity) value can have different impact (importance) based on the corresponding Nucleotide BDM value of a data point. This last observation explains why NuPoP (Affinity) does not noticeably correlate with (off-)target activity, yet is an important feature for both models, since they include NuPoP (Affinity) and Nucleotide BDM scores simultaneously.

Our results indicate that only a few out of 13 nucleosome organization-related scores show noticeable correlation with (off-)target activity or are important for model predictions. Most of these high-importance features 'measure' nucleosome affinity rather than nucleosome occupancy. Consequently, we speculate that the influence of high nucleosome affinity on Cas9 (off-)target activity exceeds that of high nucleosome occupancy. Such speculation is in concordance with the low impact of the NuPoP (Occupancy) score (see Figs. 3 and 4) on model predictions.

Conclusions

For all off-target sites featured in the crisprSQL Cas9 off-target database, we obtained 19 epigenetic features, 15 of which were newly considered. The introduced computed features characterize nucleosome organization, and include features based on BDM-based or NuPoP (Affinity). We also considered six experimental epigenetic features, namely CTCF, DNase I, DRIP, H3K4me3, MNase and RRBS. We showed that the computed features exhibited considerably larger correlation with off-target cleavage activity when compared to the six experimental epigenetic features. Interestingly, only the features CTCF, DNase I, H3K4me3 and RRBS have been frequently used in deep learning-based off-target activity prediction models. As expected, nucleosome positioning negatively impacts off-target activity. This is shown by the low Nucleotide BDM scores assigned to putative off-target sites with no detectable off-target activity. We explain this phenomenon by the presence of positioned nucleosomes which inhibit Cas9 binding. Including empirical estimates of sgRNA-DNA heteroduplex binding energies as inputs, we constructed an XGBoost tree and a CNN model. The two models were used in order to gain feature importance values of all epigenetic features. Next, we created a SHAP summary plot for each model, with feature contribution quantified by the average SHAP feature importance value across data points. The plots showed GC147, Nucleotide BDM and NuPoP (Affinity) as features among the top five which contribute most to the model's output in both models. Their importance in the two models are unlike the six experimental epigenetic scores. We uploaded the off-target cleavage activity dataset used in order to make the experimental epigenetic and computed nucleosome organization-related scores available for further research. This dataset can be found as a compressed Parquet file at https://crisprsql.com/downloads/260520_putative_nucleosomal.parquet.gz. For future work, computed scores could be combined with target sequence and binding energy features in more robust and complete CRISPR-Cas9 off-target activity prediction models. Notably, BDM-derived and NuPoP scores could be used in such models. It would also be fruitful to scrutinize whether BDM-derived and NuPoP (Affinity) are also predictive of off-target activity in other CRISPR-Cas systems.

Methods

crisprSQL

The crisprSQL database consists of experimental off-target sites and cleavage activities from 15 human CRISPR-Cas9 off-target studies. In order to conduct a comprehensive investigation on the effect of epigenetics and nucleosomes on CRISPR-Cas9 off-target activity, we

utilize crisprSQL [39]. crisprSQL is an up-to-date Cas9 off-target database containing sequence and epigenetic information for over 25,000 gRNA-off-target pairs from various human and rodent cell lines. Different experimental techniques were used to measure off-target activity in different studies. Consequently, we combine the experimental off-target cleavage activities from each study by applying a Box-Cox transformation. The transformation is such that the resulting combined cleavage activity data approximates a Gaussian with mean $= 0$ and standard deviation $= 2$, as suggested in [39]. Transformed values were clipped to the $[-4, 4]$ range, with cleavage activity values below the lowest reported assay accuracy of 10^{-5} set to -4. We furthermore augment the sites in crisprSQL with those in the respective genome which have less than seven mismatches compared to any experimental data point. These data points are assumed to have no off-target activity ($CA = -4$). Using the sequence alignment tool batmis for this [62], we generate 226,682 augmented data points. This results in a total of 251,854 data points in our dataset. In summary, the above steps yield a crisprSQL-derived dataset which was augmented with putative off-targets.

Experimental nucleosome occupancy data

The NucPosDB database [58] consists of experimental nucleosome positioning and occupancy data aggregated from various biological publications. Micrococcal Nuclease digestion with deep sequencing (MNase-seq) data are indicative of nucleosome occupancy and chromatin accessibility. In addition, MNase-seq may be indicative of CRISPR-Cas9 off-target activity. Consequently, MNase-seq data for human cell lines present in crisprSQL are extracted from NucPosDB where available. This yields three HeLa (GSM1602359 [63], GSM2680344-2680347 [64]), five K562 (GSE78984 [65], GSM920557 [66], GSM2083137-2083140 [65]) and two U2OS (GSM1838910-1838911 [67]) MNase-seq tracks. Such tracks for HeLa, K562 and U2OS are then used for annotating crisprSQL off-target sites observed in the corresponding cell line.

Adding epigenetic scores

To construct the dataset for our study, we extract the 23bp target DNA sequence and 169bp target-centered sequence context for all gRNA-target pairs in crisprSQL. We also extract the experimental epigenetic (i.e., CTCF, DNase I, H3K4me3 and RRBS) scores and the normalized off-target cleavage activity for all aforementioned gRNA-target pairs. To create a single experimental epigenetic MNase feature from the cell-specific tracks, we first average HeLa data from replicate tracks GSM2680344 and GSM2680345. Next, we average U2OS data from replicate tracks GSM1838910 and GSM1838911, and directly adopt GSM2083140 for K562. We then linearly rescale each of the three resulting sets of MNase data to [0, 1], and concatenate the sets together into a single feature. We assign zeros to off-target sites with no available MNase data. In summary, this yields a crisprSQL-derived dataset with 6 experimental epigenetic scores for each of the experimental and putative off-target sites.

Adding nucleosome organization-related scores

Various existing procedural and training-based data-driven computational tools are used for predicting nucleosome organization-related scores such as nucleosome occupancy and positioning. Whereas training-free procedural tools are adopted wherever available, only three recently developed training model-based tools, namely, NuPoP [54], nuCpos [55] and LeNup, were adopted. This is because these tools attain similar performances to the gold standard nucleosome occupancy model from Kaplan et al. [68, 69]. Alternatively, they use chemical cleavage-based nucleosome positioning data [55, 70] which have higher resolution compared to the MNase-seq data used in the gold standard model.

We further augment the crisprSQL dataset with nucleosome organization-related scores. This is done by computing nucleosome occupancy and/or positioning-related scores for each base pair in the 23bp target sequence for all off-target sites. To compute a variety of scores for each 169bp sequence context, we use a comprehensive set of nucleosome organization-related tools. The names of these tools are GC content (abbreviated GC147) [51], W/S scheme [52, 53], YR scheme [52, 53], Van Der Heijden [56], Block Decomposition Method (BDM) [48, 49], NuPoP [54], nuCpos [55], and LeNup [57]. Note that nucleosome organization-related tools like BDM [48] cannot handle 'N'-containing input sequences. As a result, the dataset used in this study only consider off-target sites with non-'N'-containing sequence contexts.

The following subsections details how each tool is used for computing one or more nucleosome organization-related scores. Since NuPoP and nuCpos both output histone affinity, nucleosome occupancy, and Viterbi scores, we include all three scores as separate features for both tools. We also derive Nucleotide BDM and Strong-Weak BDM scores from BDM. As a result, the 8 tools above generate 13 computed scores. In summary, the above steps yield a crisprSQL-derived dataset which was augmented with putative off-targets. In terms of features, it has a total of 6 experimental epigenetic and 13 nucleosome organization-related computed features. We further refer to these 19 features as epigenetic features.

GC content

GC content (or GC147 as abbreviated here for clarity) is a simple training-free measure. It is defined as the fraction of guanine and cytosine residues present in the 147bp nucleosomal sequence around a given nucleotide. Details on the use of GC content for predicting nucleosome occupancy can be found in the supplementary material.

We compute base pair-resolved GC147 values for each (off-)target site in the crisprSQL dataset. To do this, we slide a 147bp window across the (off-)target site's 169bp context sequence, thereby obtaining 23 subsequences of length 147. A GC147 value is then computed for each of these subsequences.

W/S and YR schemes

W/S and YR schemes are training-free scores used for the prediction of rotational and translational nucleosome positioning, respectively [52]. The two schemes are available on the web platform nuMap [53], and are based on sequence-dependent DNA anisotropy. Details regarding how W/S and YR schemes work can be found in the supplementary material.

We compute base pair-resolved W/S and YR Scheme values for each (off-)target site in the crisprSQL dataset. The general approach for doing this is identical to that of GC147. Namely, we slide a 147bp window across the (off-)target site's 169bp context sequence, thereby obtaining 23 subsequences of length 147. The only difference is that we use W/S and YR Scheme instead of GC147 when computing values for each of the 23 subsequences.

Van Der Heijden algorithm

In reference [56], the authors propose a method for predicting the intrinsic nucleosome position of a genome based on statistical mechanics. We abbreviate this method as VanDerHeijden. Details regarding how VanDerHeijden works can be found in the supplementary material.

We compute base pair-resolved VanDerHeijden values for all (off-)target sites in the crisprSQL dataset. To compute a VanDerHeijden score for a given (off-)target site, we first obtain the 169bp context sequence of the given site. The context sequence is then padded with 73 A nucleotides on both ends, and then passed into the Van Der Heijden algorithm (see Supplementary Material). Reading the middle 23 values in the array of 169 values produced by the algorithm then yields the base pair-resolved values. We use the following hyperparameters for VanDerHeijden:

- a nucleosome positioning window of $N = 147$,
- probability amplitude $B = 0.16$,
- dinucleotide periodicity $p = 10.1$, and
- chemical potential $\mu = -0.6$.

An implementation of the algorithm can be found at https://github.com/JvN2/NucTool.

Block decomposition method-based measures

Many recent nucleosome occupancy tools such as NuPoP are statistical and entropy-based. However, such tools often require the use of experimental nucleosome occupancy data for the training of many parameters in the model, which is computationally expensive. To resolve this, we can use the Block Decomposition Method (BDM) [48], which is a training-free method for approximating the algorithmic complexity of sequences. A consequence of this definition is that repetitive sequences, e.g., "ATATAT ATAT", have low BDM values. A recent study [49] showed that BDM scores of 147bp candidate DNA sequences carry valuable information related to nucleosome organization.

Based on BDM, we derive Nucleotide BDM, which computes the BDM of the 147bp DNA string. We also derive Strong-Weak BDM, which applies the strong-weak transformation before computing the BDM of the resulting modified string. The strong-weak transformation replaces 'G' and 'C' with 'S' (Strong) and 'A' and 'T' with 'W' (Weak) in the DNA string. We compute base pair-resolved Nucleotide BDM and Strong-Weak BDM values for each (off-)target site in the crisprSQL dataset. The general approach for doing this is identical to that of GC147. Namely, we slide a 147bp window across the (off-)target site's 169bp context sequence, thus obtaining 23 subsequences of length 147. We then use PyBDM, a Python [71] implementation of BDM, to compute Nucleotide BDM and Strong-Weak BDM values for each of the 147bp subsequences. The Python implementation of BDM can be found in https://github.com/sztal/pybdm.

NuPoP

Using a duration Hidden Markov Model (dHMM), NuPoP [54] predicts nucleosome positioning and occupancy. NuPoP accounts for the different linker length distributions or base compositions in different eukaryotes in order to make better predictions [72]. Details on NuPoP can be found in the supplementary material. An implementation of NuPoP can be found at https://github.com/jipingw/NuPoP.

We compute base pair-resolved NuPoP (Affinity), NuPoP (Occupancy) and NuPoP (Viterbi) values for each (off-)target site in the crisprSQL dataset. First, the 294,989 context sequences in the crisprSQL dataset were split into 9 sets of size 31,645 and 1 set of 10,184.

This is to accommodate the fact that NuPoP requires an input sequence length of at least 1000bp. Long strings of length $147 + (147 + 169) * 31,645 = 9,999,967$ were created for the first 9 set by adding 147 A nucleotides between each context sequence. To remove end effects, the long string also contains 147 A nucleotides both before the first context sequence and after the last context sequence. In the same way, a short string of length $147 + (147 + 169) * 10,184 = 3,218,291$ is created for the final set. The 10 long strings are then fed into the NuPoP R package individually using the `predNuPoP` function. This gives rise to 10 TSV files containing the base pair-resolved histone binding affinity, occupancy and Viterbi values. When calling `predNuPoP`, we use parameters `species=1` and `model=4`.

nuCpos

Building on NuPoP, nuCpos [55] is a recent dHMM-based algorithm for predicting nucleosome positioning. nuCpos uses the same training and inference algorithms as NuPoP. However, it improves upon NuPoP by using high-resolution H3Q85C-seq budding yeast data [70] instead of the low-resolution MNase-seq data. Similar to NuPoP, nuCpos produces histone binding affinity, predicted nucleosome occupancy and Viterbi scores. More details on the algorithm can be found in [55]. An implementation of nuCpos can be found at https://github.com/hkatomed/nuCpos.

We compute base pair-resolved nuCpos (Affinity), nuCpos (Occupancy) and nuCpos (Viterbi) values for each (off-)target site in the crisprSQL dataset. The nuCpos R package has similar input-output interfaces to NuPoP. Consequently, we use the same approach as that described for NuPoP above in order to produce these base pair-resolved values. When calling `predNuCpos`, we use parameters `species="c"`, `smoothHBA=FALSE` and `ActLikePredNuPoP=TRUE`.

LeNup

In light of the recent rise of state-of-the-art deep learning methods for data-based models, LeNup uses a convolutional neural network (CNN) with gated Inception-like modules [73, 74]. LeNup is used for nucleosome positioning prediction in a variety of eukaryotic genomes [57]. The original implementation of LeNup is available at https://github.com/biomedBit/LeNup.

LeNup was originally trained for separating nucleosomal and non-nucleosomal DNA. Consequently, we retrained the neural network used in LeNup using high resolution H3Q85C chemical cleavage-seq [70] yeast data. Because of this modification, we will refer to this measure as LeNup (H3Q85C). The retrained PyTorch [75] model can be found at https://github.com/jeffmak/crispr-cas9-epigenetics. We compute base pair-resolved LeNup (H3Q85C) values for all (off-)target site in the crisprSQL dataset. For each (off-)target site, we one-hot encode its context sequence and pass it into the PyTorch model, which outputs the base pair-resolved value.

Correlation and distribution analysis

We compute the Spearman and Pearson correlations with off-target cleavage activities for all epigenetic features. This enables us to examine the relationship between each epigenetic feature and off-target cleavage activity, and to identify features which significantly correlate with off-target activity. We also consider whether such correlations vary between gene and non-gene bodies or across cell lines. The calculation of gene bodies is not cell line dependent. The nucleosome organization-related scores are at base-pair resolution. Consequently, we take the mean of the values at each (off-) target if the score is not binary and the median of the values otherwise. Using the dataset which was augmented with putative off-targets, we separate the data points into lowest ($CA = -4$), low ($CA \leq 2$) and high ($CA > 2$) cleavage activity. We also visualize the epigenetic score distributions for these data points. In order to compare cleavage frequencies across studies, we use the nonlinear Box-Cox transformation [76] to transform cleavage rates. We transform cleavage rates to approximate a Gaussian with zero mean and standard deviation $\sigma = 2$ for each study individually. To achieve a fixed value range and treat outliers efficiently, this distribution has been clipped at -2σ and 2σ. This has been used in the literature [77, 78] before. Based on these, we separate the data points into lowest cleavage activity ($CA = -2\sigma = -4$), low cleavage activity ($CA \leq \sigma = 2$) and high cleavage activity ($CA > \sigma$).

CRISPRspec

The crisprSQL database includes estimates for the free energy of the DNA-RNA heteroduplex generated by the CRISPRspec [50] biophysical interaction model. These interaction energies are features derived from secondary structures. These features shape the thermodynamic advantage to gRNA-DNA hybrid formation upon binding of the gRNA-Cas9 complex to the off-target site. Computationally, for a given (off-)target region, CRISPRspec uses four empirical free energy contributions terms, namely:

- a PAM-dependent correcting factor δ_{PAM},
- free energy $\Delta G_H^{\text{RNA:DNA}}$ from hybridizing the gRNA and target strand, weighted by a position-wise estimate of the Cas9 influence in the binding,
- free energy $\Delta G_U^{\text{RNA:RNA}}$ from forming the secondary structure of the 20nt gRNA spacer sequence, computed using RNAFold,

- free energy $\Delta G_O^{\text{DNA:DNA}}$ from forming the dsDNA duplex from the target and non-target DNA strands.

These four terms are used for computing the total binding free energy

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

From the values given in the crisprSQL database, we calculate three key energy features to be included in our model, namely

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

Model and SHAP

CRISPR recently saw an increase in computational tools for Cas9 off-target activity prediction [79], with recent tools using machine and deep learning techniques [33, 35, 59, 80, 81]. To determine how all 19 epigenetic scores relate to off-target activity within a Cas9 off-target cleavage activity prediction model, we build two machine learning models. The first one is an extreme gradient boosted (XGBoost) tree model [82], and the second one a convolutional neural network (CNN) model. These models take three CRISPRspec-derived energy features [50], experimental epigenetic features and nucleosomal organization-related features. The CNN's model architecture is similar to DeepCRISPR's [33] Siamese neural network, but lacks the sequence arm (see Supplementary Fig. 1 for details on the architecture). Any nucleosome organization-related feature is calculated at base pair resolution leading to 23 values for an (off-)target DNA. In contrast, the mean value across the 23 (off-)target base pairs is presented for any experimental epigenetic feature.

Regarding training and evaluation for the XGBoost and CNN models, the dataset is randomly split into a training dataset and test dataset. A ratio of 80%-20% is used for the splitting. The train-test split is done in a way so as to ensure equal amounts of experimentally measured and augmented data in both datasets. For XGBoost, the tree model is trained for 70 epochs, where a new training batch with 50,000 data points is sampled in each epoch. We chose hyperparameters `eta`=0.5, `colsample_bytree`=0.7, `max_depth`=7. As for CNN, the model is trained for 70 epochs, where a new training batch with 35,000 data points is sampled in each epoch. We use hyperparameters `lr`=0.001, `batchnorm_momentum`=0.1, together with early stopping. For both models, bootstrap sampling ensures that each training batch contains equal amounts of active (CA > -4) and inactive/putative (CA $= -4$) (off-)targets. We then use the Shapley Additive Explanation (SHAP) library's Tree Explainer and Deep Explainer [83]. We use these explainers on a batch of 10,000 datapoints randomly sampled from the test data. This allows us to measure the contribution of each input feature towards the XGBoost and CNN model's prediction respectively. Contributions for each input features are then visualized using SHAP summary plots. When creating the SHAP summary plots, for each data point, we compute the SHAP contribution of each computed feature in the SHAP summary plots. The SHAP contribution for each computed feature is computed by summing up the corresponding base pair-resolved SHAP contributions.

Abbreviations

CRISPR: Clustered regularly interspaced short palindromic repeats; gRNA: Guide RNA; XGBoost: Extreme Gradient Boost; CNN: Convolutional neural network; SHAP: Shapley Additive Explanation.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-09012-7.

Additional file 1: Supplementary Table 1 Spearman and Pearson correlation values between epigenetic features and SpCas9 off-target cleavage activities. **Supplementary Figure 1** Convolutional neural network architecture used for CRISPR-Cas9 off-target activity prediction. **Supplementary Figure 2** Heatmaps showing Spearman and Pearson correlations between epigenetic features and Cas9 off-target cleavage activities for HeLa cell line data. **Supplementary Figure 3** Heatmaps showing Spearman and Pearson correlations between epigenetic features and Cas9 off-target cleavage activities for K562 and U2OS cell line data. **Supplementary Figure 4** Heatmaps showing Spearman and Pearson correlations between epigenetic features and Cas9 off-target cleavage activities for K562 and U2OS cell line data. **Supplementary Figure 5** Violin plots for all epigenetic features. **Supplementary Figure 6** Distribution plots for all epigenetic features. **Supplementary Figure 7** Heatmap showing the mean absolute value of the SHAP values for the extreme gradient boosted tree's base pair-resolved input features. **Supplementary Figure 8** Heatmap showing the mean absolute value of the SHAP values for the convolutional neural network's base pair-resolved input features. **Supplementary Figure 9** Spearman and Pearson Correlations between NuPoP (Affinity) and Nucleotide BDM across different cell lines (U2OS, HEK293, K562, HeLa) and regions (Gene Body, Not Gene Body) for the dataset used in Fig. 1. **Supplementary Figure 10** Bar plot showing Spearman and Pearson correlations between 19 epigenetic features and SpCas9 on-target cleavage activities for all cell lines that contribute more than 1% to the crisprSQL dataset. **Supplementary Figure 11** SHAP dependency plots for GC147, Nucleotide BDM and NuPoP (Affinity) for XGBoost model. **Supplementary Figure 12** SHAP dependency plots for GC147, Nucleotide BDM and NuPoP (Affinity) for CNN model.

Acknowledgements

We acknowledge careful reading of the manuscript by Shishir Rao. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558).

Authors' contributions

JM: conceptualization, data collection, analysis, writing of the article. FS: analysis, data collection, writing of the article, funding acquisition. PM:

conceptualization, analysis, writing of the article, funding acquisition. All authors have read and approved the manuscript.

Funding
Biotechnology and Biological Sciences Research Council BB/M011224 and BB/S507593/1. Funding for open access charge: Oxford University RCUK Open Access Block Grant. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Availability of data and materials
The crisprSQL database is available at http://www.crisprsql.com, where the full data set can be downloaded in CSV format. The dataset used for analysis in this paper is available as a compressed Parquet file at https://crisprsql.com/downloads/260520_putative_nucleosomal.parquet.gz, where the full data set can be downloaded in parquet format. Users are not required to log in to access any of the database features. Sample Python scripts for using the XGBoost and CNN models are available at https://github.com/jeffmak/crispr-cas9-epigenetics.

Declarations

Ethics approval and consent to participate
All human data used in this study was gained and processed in accordance with the Declaration of Helsinki. The University of Oxford Medical Sciences Interdivisional Research Ethics Committee confirms that the research conducted did not require ethics approval because it utilized previously collected anonymous data. The researchers have no access to identifying information of the people from whom the original data was obtained.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 20 April 2022 Accepted: 17 October 2022
Published online: 06 December 2022

References

1. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007;315(5819):1709–12. https://doi.org/10.1126/science.1138140.
2. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science. 2012;337(6096):816–21. https://doi.org/10.1126/science.1225829.
3. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science. 2013;339(6121):819–23. https://doi.org/10.1126/science.1231143.
4. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol. 2014;32(4):347–55.
5. Tsai SQ, Joung JK. Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. Nat Rev Genet. 2016;17(5):300–12.
6. Adli M. The CRISPR tool kit for genome editing and beyond. Nat Commun. 2018;9(1):1911.
7. Zhang F. Development of CRISPR-Cas systems for genome editing and beyond. Q Rev Biophys. 2019;52:e6. https://doi.org/10.1017/S0033583519000052.
8. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013;152(5):1173–83.
9. Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK. CRISPR RNA-guided activation of endogenous human genes. Nat Methods. 2013;10(10):977–9.
10. Perez-Pinera P, Kocak DD, Vockley CM, Adler AF, Kabadi AM, Polstein LR, et al. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Methods. 2013;10(10):973–6.
11. Ma H, Naseri A, Reyes-Gutierrez P, Wolfe SA, Zhang S, Pederson T. Multicolor CRISPR labeling of chromosomal loci in human cells. Proceedings of the National Academy of Sciences. 2015;112(10):3002–7. https://doi.org/10.1073/pnas.1420024112.
12. Shao S, Zhang W, Hu H, Xue B, Qin J, Sun C, et al. Long-term dual-color tracking of genomic loci by modified sgRNAs of the CRISPR/Cas9 system. Nucleic Acids Res. 2016;44(9):e86.
13. Kearns NA, Pham H, Tabak B, Genga RM, Silverstein NJ, Garber M, et al. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nat Methods. 2015;12(5):401–3.
14. Kwon DY, Zhao YT, Lamonica JM, Zhou Z. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. Nat Commun. 2017;8:15315.
15. Wang H, Xu X, Nguyen CM, Liu Y, Gao Y, Lin X, et al. CRISPR-Mediated Programmable 3D Genome Positioning and Nuclear Organization. Cell. 2018;175(5):1405-1417.e14.
16. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013;31(9):822–6.
17. Cradick TJ, Fine EJ, Antico CJ, Bao G. CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity. Nucleic Acids Res. 2013;41(20):9584–92.
18. Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res. 2014;42(11):7473–85.
19. Guilinger JP, Pattanayak V, Reyon D, Tsai SQ, Sander JD, Joung JK, et al. Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. Nat Methods. 2014;11(4):429–35.
20. Fujita T, Yuno M, Fujii H. Allele-specific locus binding and genome editing by CRISPR at the p16INK4a locus. Sci Rep. 2016;6:30485.
21. Kallimasioti-Pazi EM, Thelakkad Chathoth K, Taylor GC, Meynert A, Ballinger T, Kelder MJE, et al. Heterochromatin delays CRISPR-Cas9 mutagenesis but does not influence the outcome of mutagenic DNA repair. PLOS Biol. 2018;16(12):1–22. https://doi.org/10.1371/journal.pbio.2005595.
22. O'Geen H, Henry IM, Bhakta MS, Meckler JF, Segal DJ. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. Nucleic Acids Res. 2015;43(6):3389–404.
23. Horlbeck MA, Witkowsky LB, Guglielmi B, Replogle JM, Gilbert LA, Villalta JE, et al. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. eLife. 2016;5:e12677.
24. Kuscu C, Arslan S, Singh R, Thorpe J, Adli M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat Biotechnol. 2014;32(7):677–83.
25. Chen Y, Zeng S, Hu R, Wang X, Huang W, Liu J, et al. Using local chromatin structure to improve CRISPR/Cas9 efficiency in zebrafish. PLoS ONE. 2017;12(8):1–19. https://doi.org/10.1371/journal.pone.0182528.
26. Jensen KT, Fløe L, Petersen TS, Huang J, Xu F, Bolund L, et al. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. FEBS Lett. 2017;591(13):1892–901.
27. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010;2010(2):pdb.prot5384.
28. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18):5868–77. https://doi.org/10.1093/nar/gki901.
29. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011;6(4):468–81.
30. O'Geen H, Echipare L, Farnham PJ. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. Methods Mol Biol. 2011;791:265–86.
31. Verkuijl SA, Rots MG. The influence of eukaryotic chromatin state on CRISPR-Cas9 editing efficiencies. Curr Opin Biotechnol. 2019;55:68–73.
32. Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol. 2014;32(7):670–6.
33. Chuai G, Ma H, Yan J, Chen M, Hong N, Xue D, et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol. 2018;19(1):80.
34. Liu Q, Cheng X, Liu G, Li B, Liu X. Deep learning improves the ability of sgRNA off-target propensity prediction. BMC Bioinformatics. 2020;21(1):51.

35. Liu Q, He D, Xie L. Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas System using attention boosted deep learning and network-based gene feature. PLOS Comput Biol. 2019;15(10):1–22. https://doi.org/10.1371/journal.pcbi.1007480.
36. Kim S, Yu NK, Kaang BK. CTCF as a multifunctional protein in genome regulation and gene expression. Exp Mol Med. 2015;47:e166.
37. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22(9):1813–31.
38. Liu X, Wang C, Liu W, Li J, Li C, Kou X, et al. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. Nature. 2016;537(7621):558–62.
39. Störtz F, Minary P. crisprSQL: a novel database platform for CRISPR/Cas off-target cleavage assays. Nucleic Acids Res. 2021;49(D1):D855–61. https://doi.org/10.1093/nar/gkaa885.
40. Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell. 2012;45(6):814–25.
41. Ginno PA, Lim YW, Lott PL, Korf I, Chédin F. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res. 2013;23(10):1590–600.
42. Al-Hadid Q, Yang Y. R-loop: an emerging regulator of chromatin dynamics. Acta Biochim Biophys Sin (Shanghai). 2016;48(7):623–31.
43. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013;20(3):267–73. https://doi.org/10.1038/nsmb.2506.
44. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008;132(5):887–98.
45. Kuan PF, Huebert D, Gasch A, Keles S. A non-homogeneous hidden-state model on first order differences for automatic detection of nucleosome positions. Stat Appl Genet Mol Biol. 2009;8:Article29.
46. Hinz JM, Laughery MF, Wyrick JJ. Nucleosomes Inhibit Cas9 Endonuclease Activity in Vitro. Biochemistry. 2015;54(48):7063–6.
47. Isaac RS, Jiang F, Doudna JA, Lim WA, Narlikar GJ, Almeida R. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. eLife. 2016;5:e13450.
48. Zenil H, Hernández-Orozco S, Kiani NA, Soler-Toscano F, Rueda-Toicen A. A Decomposition Method for Global Evaluation of Shannon Entropy and Local Estimations of Algorithmic Complexity. 2016. arXiv:1609.00110
49. Zenil H, Minary P. Training-free measures based on algorithmic probability identify high nucleosome occupancy in DNA sequences. Nucleic Acids Res. 2019;47(20):e129–e129. https://doi.org/10.1093/nar/gkz750.
50. Alkan F, Wenzel A, Anthon C, Havgaard JH, Gorodkin J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. Genome Biol. 2018;19(1):177.
51. Tillo D, Hughes TR. G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics. 2009;10(1). https://doi.org/10.1186/1471-2105-10-442.
52. Cui F, Zhurkin VB. Structure-based Analysis of DNA Sequence Patterns Guiding Nucleosome Positioningin vitro. J Biomol Struct Dyn. 2010;27(6):821–41. https://doi.org/10.1080/073911010010524947.
53. Alharbi BA, Alshammari TH, Felton NL, Zhurkin VB, Cui F. nuMap: A Web Platform for Accurate Prediction of Nucleosome Positioning. Genomics Proteomics Bioinforma. 2014;12(5):249–53. https://doi.org/10.1016/j.gpb.2014.08.001.
54. Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP. Predicting nucleosome positioning using a duration Hidden Markov Model. BMC Bioinformatics. 2010;11(1):346. https://doi.org/10.1186/1471-2105-11-346.
55. Kato H, Shimizu M, Urano T. Chemical map–based prediction of nucleosome positioning using the Bioconductor package nuCpos. bioRxiv. 2019. https://doi.org/10.1101/2019.12.25.888305.
56. van der Heijden T, van Vugt JJFA, Logie C, van Noort J. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. Proc Natl Acad Sci. 2012;109(38):E2514–22. https://doi.org/10.1073/pnas.1205659109.
57. Zhang J, Peng W, Wang L. LeNup: learning nucleosome positioning from DNA sequences with improved convolutional neural networks. Bioinformatics. 2018;34(10):1705–12. https://doi.org/10.1093/bioinformatics/bty003.
58. Shtumpf M, Piroeva KV, Agrawal SP, Jacob DR, Teif VB. NucPosDB: a database of nucleosome positioning in vivo and nucleosomics of cell-free DNA. Chromosoma. 2022. https://doi.org/10.1007/s00412-021-00766-9.
59. Zhang G, Dai Z, Dai X. C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. Comput Struct Biotechnol J. 2020;18:344–54.
60. Cofsky JC, Soczek KM, Knott GJ, Nogales E, Doudna JA. CRISPR-Cas9 bends and twists DNA to read its sequence. Nat Struct Mol Biol. 2022;29(4):395–402.
61. Vinogradov AE. DNA helix: the importance of being GC-rich. Nucleic Acids Res. 2003;31(7):1838–44.
62. Tennakoon C, Purbojati RW, Sung WK. BatMis: a fast algorithm for k-mismatch mapping. Bioinformatics. 2012;28:2122–8.
63. Kfir N, Lev-Maor G, Glaich O, Alajem A, Datta A, Sze SK, et al. SF3B1 association with chromatin determines splicing outcomes. Cell Rep. 2015;11(4):618–29.
64. Schwartz U, Németh A, Diermeier S, Exler JH, Hansch S, Maldonado R, et al. Characterizing the nuclease accessibility of DNA in human cells to map higher order structures of chromatin. Nucleic Acids Res. 2018;47(3):1239–54. https://doi.org/10.1093/nar/gky1203.
65. Mieczkowski J, Cook A, Bowman SK, Mueller B, Alver BH, Kundu S, et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. Nat Commun. 2016;7:11485.
66. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res. 2012;22(9):1735–47.
67. Devaiah BN, Case-Borden C, Gegonne A, Hsu CH, Chen Q, Meerzaman D, et al. BRD4 is a histone acetyltransferase that evicts nucleosomes from chromatin. Nat Struct Mol Biol. 2016;23(6):540–8.
68. Liu H, Zhang R, Xiong W, Guan J, Zhuang Z, Zhou S. A comparative evaluation on prediction methods of nucleosome positioning. Brief Bioinform. 2013;15(6):1014–27. https://doi.org/10.1093/bib/bbt062.
69. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2008;458(7236):362–6. https://doi.org/10.1038/nature07667.
70. Chereji RV, Ramachandran S, Bryson TD, Henikoff S. Precise genome-wide mapping of single nucleosomes and linkers in vivo. Genome Biol. 2018;19(1). https://doi.org/10.1186/s13059-018-1398-0.
71. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley: CreateSpace; 2009.
72. Burshtein D. Robust parametric modeling of durations in hidden Markov models. IEEE Trans Speech Audio Process. 1996;4(3):240–2.
73. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. CoRR. 2015. arXiv:1512.00567.
74. Dauphin YN, Fan A, Auli M, Grangier D. Language Modeling with Gated Convolutional Networks. CoRR. 2016. arXiv:1612.08083.
75. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Red Hook: Curran Associates, Inc.; 2019. p. 8024–35.
76. Box GE, Cox DR. An analysis of transformations. J R Stat Soc Ser B (Methodol). 1964;26(2):211–43.
77. Listgarten J, Weinstein M, Kleinstiver BP, Sousa AA, Joung JK, Crawford J, et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat Biomed Eng. 2018;2(1):38–47.
78. Wang J, Xiang X, Bolund L, Zhang X, Cheng L, Luo Y. GNL-Scorer: a generalized model for predicting CRISPR on-target activity by machine learning and featurization. J Mol Cell Biol. 2020;12(11):909–11. https://doi.org/10.1093/jmcb/mjz116.
79. Bradford J, Perrin D. A benchmark of computational CRISPR-Cas9 guide design methods. PLoS Comput Biol. 2019;15(8):e1007274.
80. Haeussler M, Schönig K, Eckert H, Eschstruth A, Mianné J, Renaud JB, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. 2016;17(1):148.
81. Charlier J, Nadon R, Makarenkov V. Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-Cas9 gene editing. Bioinformatics. 2021;Btab112. https://doi.org/10.1093/bioinformatics/btab112.

82. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 785-794. https://doi.org/10.1145/2939672.2939785.
83. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook: Curran Associates Inc.; 2017. p. 4768–77.

Publisher's Note