biomolecules

Article

A Machine Learning Approach to Identify the Importance of Novel Features for CRISPR/Cas9 Activity Prediction

Dhvani Sandip Vora [1], Yugesh Verma [1] and Durai Sundar [1,2,*]

[1] Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[2] Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[*] Correspondence: sundar@dbeb.iitd.ac.in

Abstract: The reprogrammable CRISPR/Cas9 genome editing tool's growing popularity is hindered by unwanted off-target effects. Efforts have been directed toward designing efficient guide RNAs as well as identifying potential off-target threats, yet factors that determine efficiency and off-target activity remain obscure. Based on sequence features, previous machine learning models performed poorly on new datasets, thus there is a need for the incorporation of novel features. The binding energy estimation of the gRNA-DNA hybrid as well as the Cas9-gRNA-DNA hybrid allowed generating better performing machine learning models for the prediction of Cas9 activity. The analysis of feature contribution towards the model output on a limited dataset indicated that energy features played a determining role along with the sequence features. The binding energy features proved essential for the prediction of on-target activity and off-target sites. The plateau, in the performance on unseen datasets, of current machine learning models could be overcome by incorporating novel features, such as binding energy, among others. The models are provided on GitHub (GitHub Inc., San Francisco, CA, USA).

Keywords: CRISPR/Cas9; genome editing; machine learning; SHAP values; binding energy; off-targets

Citation: Vora, D.S.; Verma, Y.; Sundar, D. A Machine Learning Approach to Identify the Importance of Novel Features for CRISPR/Cas9 Activity Prediction. *Biomolecules* **2022**, *12*, 1123. https://doi.org/10.3390/biom12081123

Academic Editors: Cameron Mura and Lei Xie

Received: 13 July 2022
Accepted: 10 August 2022
Published: 16 August 2022

1. Introduction

Clustered regularly interspersed short palindromic repeats (CRISPR) and its associated nuclease Cas9 constitute a versatile and reprogrammable genome editing mechanism that has been repurposed as a widely used tool [1–3]. The single guide RNA can be customised to target the DNA at any location by changing the 20 nucleotides "spacer". This spacer is designed to complement the "protospacer" region in the DNA, at which the Cas9 nuclease would create a double-stranded break [4]. A 3-nucleotide protospacer adjacent motif (PAM) is a prerequisite for probing and cleaving the target DNA by this two-component protein–RNA system [1]. The PAM site is generally of the form of NGG (where N is any nucleotide) for the *Streptococcus pyogenes*-derived Cas9 (SpCas9) protein [5,6]. The SpCas9 is a multidomain protein consisting of (i) three recognition domains that bind to the RNA and DNA strands, (ii) two nuclease domains to cleave each of the DNA strands, (iii) a PAM interaction domain, and (iv) an arginine-rich helix which acts as a linker [7]. Although this system is a facile and flexible genome editing tool, there are two critical design problems associated with this system: (i) designing a guide RNA with good activity at the intended target region and (ii) ensuring that the selected guide does not show activity at similar unintended sites, or in other words, has low off-target activity [8,9]. The presence of the Cas9 off-target activity has hindered clinical applications of Cas9, which is a significant area of focus for CRISPR/Cas9 study.

Great strides have been taken to understand the mechanism of action and, consequently, develop design rules to aid experimentalists in optimising guides for the intended applications. The field has benefited greatly over the past decade, majorly because of the development of multiple methods to detect Cas9 off-target activity in vitro and in situ within the cell [10–15]. Off-target detection techniques have enabled the identification of empirical rules that seem to drive off-target identification and activity by allowing analyses of various off-targets generated for multiple guides under different conditions [16–19].

The availability of an experimentally derived structure and sequence of target and off-target data has allowed computational studies to understand Cas9 activity. Many prediction algorithms have been proposed to achieve each of the tasks mentioned above, qualitative algorithms and scoring schemes to rank guides by on-target efficiency and off-target predictions [20,21]. Most algorithms are based on sequence features—number and position of mismatches (PAM proximal ends are less likely to tolerate mismatches, while the distal ends report more tolerance for mismatches) [17]. Many machine learning models have been built to predict the performance of guides and the prediction of their respective off-targets based on rules depending on the system's various sequence and structural features [17,22–25], yet there is a gap between the predictions and experimentally observed results. Popular machine learning models are based on features such as the sequence at the cut-site, the number of mismatches, experimentally validated efficiency and off-target activity of the guides. Recently, deep learning models have been reported, which are trained on large-scale datasets, and some have included novel features for validation; for example, DeepCRISPR, one of the earlier attempts at building a deep learning-based tool for prediction, introduced four epigenetic features apart from the sequence features [26]. DeepCpf1 is a convolution neural net (CNN) model, and CRISPcut is a rule-based model, both of which include chromatin accessibility as an additional feature to improve the prediction confidence [27,28]. CRISPcut and AttnToCrispr are prediction algorithms that also have included the cell-line information as features while predicting off-targets and on-target efficiency, respectively [28,29]. The addition of new and important features has, in each case, improved the model performance and confidence in the predictions. Recent studies have reported that DNA enthalpy (a proxy for the stability of the DNA duplex) and DNA-RNA duplex energy parameters play an essential role in predicting on-target efficiency and off-target activity [24,30]. This study presents two new features that prove to be important in future prediction algorithm designs: MMGBSA-based binding energy for (i) DNA and guide RNA, and (ii) Cas9 protein–nucleic acid recognition domain and the DNA-RNA hybrid.

2. Materials and Methods

2.1. Data Assembly

The data used for model training and validation were obtained from published methods of CRISPR/Cas9 off-target site prediction (CRISPcut) [28] and detection (CIRCLE-seq) [11,28] (SRA identifier SRP103697). The predictions obtained from CRISPcut, run with default parameters, for the 11 guide RNAs used in CIRCLE-seq were used to obtain a comprehensive list of potential off-target sites in the genome for the corresponding cell lines used in the CIRCLE-seq experiment. The experimentally validated off-target sites were called the positive dataset, while the predictions not validated experimentally were referred to as the negative dataset. All predictions obtained from CRISPcut were analysed for chromatin accessibility; only accessible sequences were selected since earlier studies have established the importance of this feature [31–33]. The data assembly and selection are summarised in Table S3. The cleavage efficiency obtained from the CIRCLE-seq dataset for all reported off-targets was normalised to fit a uniform scale. The features used for model training are detailed in Table S4.

2.2. Predictive Features

Multiple predictive features were calculated for each of the sequences—mismatch position, number of mismatches, mismatch in PAM, type of mismatch (transition, transversion or indel), cell line information, percentage GC for the protospacer, percentage GC in the seed region, chromosome number, DNA strand information and the two new proposed binding energy features. Two MMGBSA-based binding energy features were considered—

dG(REC3:hybrid) and dG(DNA:RNA). The dG(REC3:hybrid) was calculated between the REC3 domain of SpCas9 and the 20-nucleotide DNA-RNA hybrid. The binding energy of the 20-nucleotide RNA and target DNA strands was calculated as dG(DNA:RNA). The MMGBSA calculations were carried out using the Schrödinger Maestro suite's Prime utility after pre-processing and the restrained minimisation of the complexes [34,35].

2.3. MMGBSA Binding Energy Calculation

The structure used as a template was obtained from RCSB PDB (ID: 4UN3). The REC3 domain was selected (residues 447–718) along with the 20 nucleotides of the target DNA and the 20 nucleotides of the guide RNA. The PyMOL nucleic acid mutagenesis tool was used to create all target and off-target systems from the template [36]. The structures were imported in the Schrödinger Maestro suite and preprocessed, hydrogen bonds were optimised, and restrained minimisation was carried out before performing MMGBSA calculation using the Prime utility [34,37]. The energies of molecular mechanics when combined with the generalized Born and surface area continuum solvation (MMGBSA) is a popular approach to estimate the binding free energy between biomolecules. MMGBSA is an intermediate in both computational costs and accuracy, widely applied for various systems [38–40]. The free energy is calculated and summed over solvation energy, gas-phase energy and entropic contributions. The REC3 domain was chosen as the receptor and the DNA-RNA hybrid was used as the ligand for the dG(REC3:hybrid) feature; DNA was selected as the receptor for the dG(DNA:RNA) feature.

2.4. Mann–Whitney U Test

The Mann–Whitney U test, also called the Mann–Whitney–Wilcoxon test, is a non-parametric test to compare differences of a variable between two groups when the variable in question is not normally distributed. The test was performed on the dataset for both dG features, the values of which served as input for the test enabled by the Pingouin Python package (0.5.2) [41]. The common language effect size was calculated using a Python script. The output is a U statistic and *p*-value, which indicates whether the groups show stochastic equality or not. The test is also robust to outliers. The U test was used to determine if the dG values for the experimentally validated off-targets (positive) and the non-validated predictions (negative) were statistically different.

2.5. Machine Learning Model Implementation

Two machine learning models were implemented:

(1) A random forest regression model on a small fraction of the CIRCLE-seq dataset with the dependant variable as normalised cleavage frequencies following their normalisation;
(2) A random forest classification model on a fraction of the CIRCLE-seq and CRISPcut derived datasets with the dependant variable being whether the sequence is cleaved experimentally or not.

The regression model was to determine whether the binding energy features significantly impact the cleavage frequency of the off-target sequences. The classification model would help determine if the energy features play a role in differentiating experimentally unlikely predictions from experimentally validated off-target sequences. Since the dG values calculation was computationally intensive and time consuming, the dataset consisted of 186 positive examples and 126 negative examples. However, the sequences were collected manually to ensure sufficient diversity in cleavage frequency, the number of mismatches, and other sequence features that were previously reported as significant. The classification model was implemented to understand if the features were sufficient to differentiate between experimentally likely predictions and those that are not.

Multiple machine learning models were tested with varying parameters; the best performing models were reported. All models evaluated were implemented using the scikit-learn package in Python [42].

2.6. Sampling Data for Training

Initial training was performed on a 75% train set, and assessment of the model performance was measured on the 25% held-out test dataset. The best performing model architecture was selected. For analysis of feature importance, since the dataset was limited, training was carried out again with 5-fold cross validation to ensure that the unbalanced dataset was not a limiting factor for model performance. The 5-fold cross-validation was repeated to ensure the absence of bias for both models.

2.7. Assessing Model Performance

The regression model's performance was evaluated by comparing the mean squared error (MSE), mean absolute error (MAE) and the R-squared values, and the better performing model was selected for feature importance determination and feature ranking. The MAE and MSE measure the difference between the model predictions and actual observations; hence, the ideal score is 0. The R-squared value is a correlation coefficient measuring a linear correlation between two continuous variables. The variance weighted measure is an explanation of the variance in the model output, the best score being 1.

The classification model was assessed using its confusion matrix:

$$\mathrm{M} = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

where *TP* stands for true positive, *FP* for false positive, *FN* for false negative and *TN* for true negative. The accuracy of a model is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The recall is the measure of how many actual positives the model can capture, while the precision is how many of the predicted positives are correct. The precision–recall curve, a standard evaluation criterion for a classification model, is based on the following definitions:

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

The F1 score, or F-measure, is the harmonic mean of the precision and recall, conveying a balance between the two. It is defined as

$$\text{F1 score} = \frac{2 * Recall * Precision}{Recall + Precision}$$

2.8. Identifying Feature Importance

Interpreting the features that impact a machine learning model's outcome is important for enabling the predictions' validation. In the regression and classification models used, the feature set is small, and so is the dataset; hence, each feature's influence must be understood. Hence, Shapley additive explanations (SHAP) values were implemented using the shap library in Python [43]; the TreeExplainer utility was used to analyse the random forest regressor output and to describe the model output of the random forest classifier [44]. The shap method employs an explanatory model with feature weights to explain relative feature importance and is adapted from game theory. It is to be noted that shap values do not indicate causality.

3. Results

3.1. Data Assembly and Processing

The guide RNAs and their respective off-targets were obtained from the CIRCLE-seq data [11]. The data obtained from the prediction algorithm CRISPcut were checked for the number of sites predicted for each guide RNA input [28]. The number of sites predicted hold little correlation with the experimental sites (Figure S1a). However, when the chromatin accessible sites were selected and compared, a sufficient correlation was obtained between the number of sites predicted and the number of sites confirmed experimentally (Figure S1b). Moreover, since chromatin accessibility has been shown in earlier studies to be an important feature, sequences selected for the model were only from the accessible sequences' subset [31,45].

The sequences selected from the CIRCLE-seq (positive dataset) and CRISPcut predictions, but not found in the experimentally validated datasets (negative dataset), were selected manually to ensure that the other features, such as the number of mismatches, cleavage frequencies and cell lines, were sufficiently represented. The features included for the model prediction were calculated using Python scripts, except the binding energy features, which were calculated using the method described. The resulting dataset had 40 features and 312 data points.

To determine if the features were correlated with each other, correlation analysis was carried out, and the results are shown in Figure S2. No significant correlation between the features was observed. The correlation islands observed were between the cell lines that were one-hot encoded and are hence mutually exclusive. A high correlation was expected for the total mismatches and protospacer mismatches (referred to as number of mismatches, #mm); the same can be stated for total PAM mismatches and types of PAM mismatches—transversion or transition type. Hence, the features selected were unique and not redundant.

3.2. Statistical Analysis of the Binding Energy Features

To determine if the values of the binding energies, by themselves, could be used to differentiate between the positive and negative datasets, the Mann–Whitney U test was carried out to compare the values between the two sets (Supplementary Table S1). The Mann–Whitney U test is a non-parametric test to check if a feature's values are larger for one of the two populations being compared; it is the non-parametric equivalent of the unpaired *t* test.

The values of the two binding energy features were compared for the positive and negative datasets, where the H_0 hypothesis was that the values for the two groups are equal. Hence, the H_0 hypothesis's rejection indicated that the difference between randomly selected values of the features from both populations is big enough to be statistically significant (Table S1). The rank–biserial correlation coefficient indicated the difference between total amount of favourable and unfavourable evidence. The common language effect size is the probability that a random value from Group 1 is greater than a random value from Group 2.

The Mann–Whitney U (MWU) test indicated that the values of the two binding energy features—dG(REC3:hybrid) and dG(DNA:RNA)—have differing values for the positive and negative datasets (Table S1). Moreover, it is evident from the MWU test that a random value from the negative dataset is likely to be higher than a random value from the positive dataset. However, since the effect size values are low, the features cannot solely be used as a distinguishing factor for the negative and positive datasets. The difference in population means the calculation was not enough to reliably call these features distinguishing.

3.3. Regression Model Selection and Performance Assessment

Linear, quadratic, cubic, multi-layer perceptrons and random forest regressors were implemented with varying parameters and random states to determine the best performing model. The dependent variable was the cleavage frequency for the off-target sequences obtained from the CIRCLE-seq dataset. The performance measured in the *R*-squared value, mean absolute error, mean squared error and variance-weighted measure is summarised in Table 1. The random forest regressor was chosen based on its superior performance on the dataset, compared to the other models tested. The random forest algorithm is known for its ability to predict well on tabular data, as is the case here. The perceptron was also tested for multiple nodes in one and two hidden layers trained till convergence; however, it failed to outperform the random forest regressor.

Table 1. Summary of model performances. All values shown are for the test dataset.

Metrics	Regressor							
	Linear	**Quadratic**	**Cubic**	**Decision Tree**	**SVR**	**MLP**	**XGBoost**	**Random Forest**
Mean Absolute Error	0.19	1.23	0.51	0.21	0.19	0.19	0.17	0.06
Mean Squared Error	0.07	3.49	0.54	0.09	0.08	0.07	0.06	0.01
Root MSE	0.26	1.87	0.73	0.29	0.28	0.26	0.24	0.08
R-squared value	0.37	−32.26	−4.15	0.18	0.27	0.42	0.47	0.94
Variance weighted	0.38	−31.82	−4.13	0.24	0.27	0.47	0.55	0.94

The various model metrics listed in the first column are given for the regression models tested. For the random forest regressor, the metrics are comparatively much better than the other three. It was selected for feature importance analysis. SVR stands for support vector regressor. MSE stands for mean squared error. The values reported for each regressor is after the optimisation of individual models.

The best performing regression model, the random forest regressor, was initialized on various random states and number of trees (as shown in Figure S3). The model with the maximum R-squared and minimum mean absolute error (MAE) was selected for further analysis, following which 5-fold cross-validation was performed. The resulting mean squared error (MSE) remained at 0.05, standard deviation (STD) was 0.01, and the R^2 score was 0.92, indicating that the chosen model was robust.

3.4. Explaining Feature Importance for the Random Forest Regressor

The importance and magnitude of the impact of the features on the model output were explored in detail since the aim of the study was to establish the importance of the two features proposed, namely the energy of binding of the REC3 domain of Cas9 to the 20 nucleotide hybrid of the target DNA and guide RNA-dG(REC3:hybrid), and the binding energy of the 20 nucleotide DNA to the guide RNA strand-dG(DNA:RNA). The variable importance plot (Figure 1a) generated by implementing SHAP [43,44,46] lists the most important features in descending order. The ones on top contributed the most to the model output and hence, have high predictive capability.

The SHAP values also help determine the relationship of the features to the output. The SHAP variable importance plot (Figure 1b) ranked variables in descending order of importance, and the horizontal spread indicated the effect of the value and the corresponding higher or lower prediction. Each dot is a value for an instance in the data, and the colour indicates a higher or lower value for that instance. While distance (total mismatches in the sequence) and #mm (mismatches in the protospacer region) were redundant features and showed a similar impact on output, Figure 1 shows that the low binding energy of the DNA-RNA hybrid, dG(DNA:RNA), had a high impact on model output; while the binding energy of the Cas9 REC3 domain to the DNA-RNA hybrid, dG(REC3:hybrid) was negatively correlated with the model output. Figure 1 also indicates that the presence of mismatch at the 6th position played an important role in determining the model output.

The SHAP variable importance plot (Figure 2) takes three values: a base value, SHAP values, and the matrix of feature values. The base value was the average or expected model output, and the SHAP value of a feature and the value of the feature at that instance determined in which direction the features “push” the model output. The output value highlighted is the model output for this instance. The features in red direct the output higher, while those in blue push the predictions lower. The SHAP plot for three instances are shown; since each feature plays a different role for each instance, it is essential to consider the local as well as global relevance of the feature.

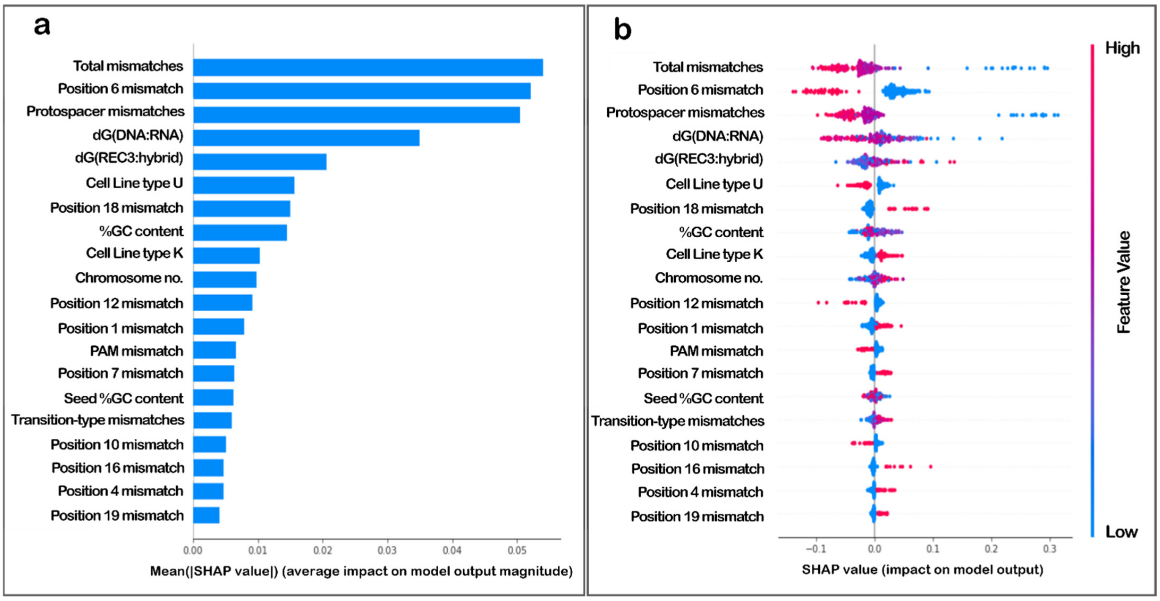

Figure 1. SHAP variable importance plots. (**a**) The plot arranges features in decreasing order of magnitude of impact on model output. (**b**) The features are listed in decreasing order of importance, the dots are coloured according to value (in a gradient from high to low, as red to blue) and the impact for each instance is plotted horizontally. The spread indicates impact on model output, and the colour indicates feature value for that output.

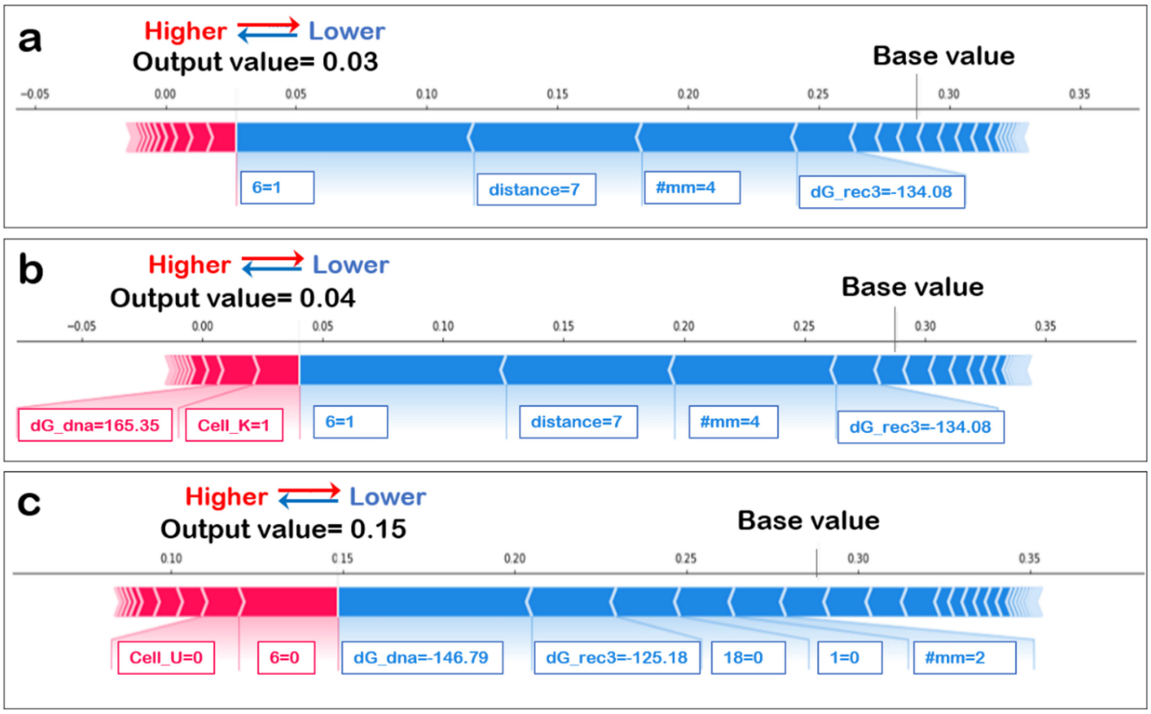

Figure 2. SHAP variable importance plot. The SHAP variable impact on outcome for singular datapoints are shown. Examples shown are explainer plots for dataset indices (**a**) 0, (**b**) 1 and (**c**) 2. The base value labelled in the figure in influenced by varying degrees by the features shown in the diagrams and the output value (shown in bold) was obtained. The features SHAP values are written alongside the features if it causes an increase in base value it is shown in red otherwise in blue.

The SHAP dependence plot (Figure 3) describes partial dependence between a feature selected, and the reference feature was chosen automatically by the script with which the chosen feature interacts the most. The dots mark each instance of the chosen variable, and the colour of the dots indicate the value of the reference feature for that instance. In Figure 3a,b, there is no clear trend between the two features; however, in Figure 3a the absence of a mismatch at position 4 and the lower values of dG(DNA:RNA) have a higher impact on the model output. Figure 3b shows that the partial dependence between the two features is not significant and no trend can be observed. The spread of the plot indicates the relationship between the two features. As in Figure 3c, the vertical dispersion at a particular value shows the interaction effect between the two features. Moreover, an approximately negative correlation exists between the variables, and a smaller Hamming distance (total mismatches in the off-target) would have more influence on the model output; it also corresponds with lower values of dG(DNA:RNA).

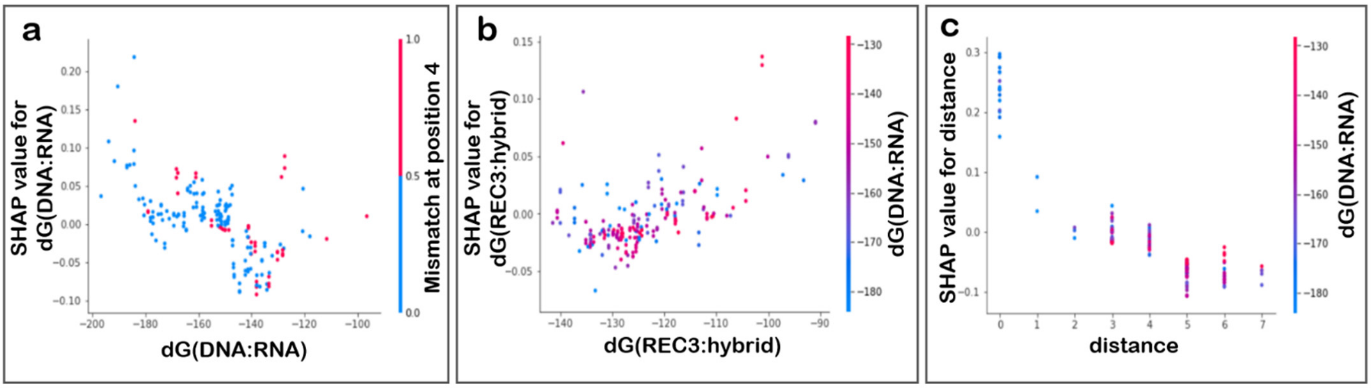

Figure 3. SHAP feature dependence plot. The plots show dependence between (**a**) dG(DNA:RNA) and a mismatch at position 4, (**b**) dG(REC3:hybrid) and dG(DNA:RNA) and (**c**) distance and dG(DNA:RNA). The vertical axis marks the SHAP values for the chosen feature, while the horizontal axis shows spread of the values of the feature. The reference feature was selected by the algorithms automatically and was used to colour the dots that indicate value of the primary feature for an instance. No clear trend can be observed in (**a**,**b**). In (**c**), vertical clusters at individual values indicate a correlation with dG(DNA:RNA) values, and the plot also shows a negative correlation of the values of the distance with the output variable.

3.5. Classifier Model Selection and Performance Assessment

The classifier models were built to study the contribution of the binding energy features to machine learning models that can distinguish between positive (sequences that are off-target sites in experiments) and negative datasets (sequences predicted to be off-targets but were not found in experiments). Various classification models were trained on the dataset, optimised for each type of model (the best performing model's accuracy summarised in Table S2). Since the random forest classifier performed well on the 25–75 test-train split, the model was evaluated after 5-fold cross validation. The classifier yielded good accuracy and was implemented for further analysis. The model metrics for the random forest classifier model are summarised in Table 2.

Table 2. Model performance of the random forest classifier, measured on test dataset.

Model Metrics	Score on Test Data	Overall Score
Accuracy	0.86	0.97
Precision	0.88	0.98
Recall	0.94	0.96
F1 score	0.91	0.97

The accuracy, precision, recall and F1 scores are calculated as mentioned in the Methods section. The accuracy reported is after 5-fold cross validation. The overall score is for combined test and train datasets.

The performance of the random forest classifier was tested using various parameters as shown in Figure 4. The model predicted the correct classes for each label reliably. The precision–recall curve and receiver operating characteristic (ROC) cover over 95% area under the curve, indicating a robust classification model. The next best performing model (support vector machine classifier) did not perform better, even on 5-fold cross validation, and hence was not evaluated further. Since the study aimed not to build an off-target determination model, but rather discern the importance of energy features, more complex models were not tested.

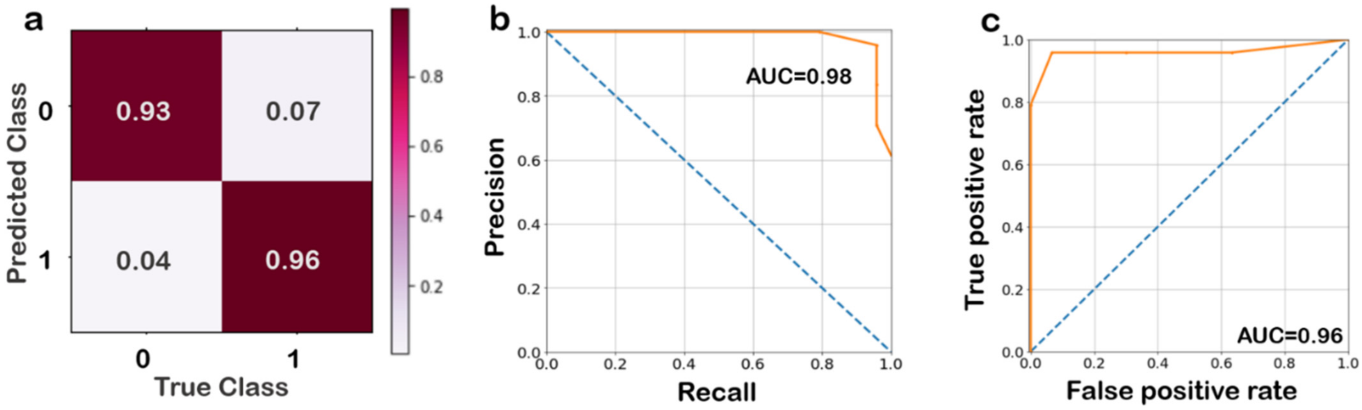

Figure 4. (**a**) Confusion matrix for the random forest classifier, vertical axis is for predicted labels and the horizontal axis states the true labels. The values are ratios of the number of instances predicted to the total instances in the class. (**b**) Precision–recall curve, shown in orange which has an area under the curve of 0.98 for the whole dataset, (**c**) receiver operating characteristic (ROC) also shown in orange for the test dataset, which plots the true positive rate against the false positive rate. The area under the curve (AUC) is 0.96. The dashed blue line across the diagonal shows 50% accuracy.

3.6. Explaining Feature Importance for the Classifier

The importance of the features in a well-performing classification model that can learn the difference between the positive and negative datasets will determine if the binding energy features play a significant role in determining the model output. The SHAP value plots for each instance are not shown for lack of space, but three examples are shown in Figure 5. The base value, determined as the average from the training dataset, is influenced by the features listed in order of magnitude of impact. Features in blue lower the output, while features in red increase the output. In all instances, energy features play an important role. However, since feature importance for each datapoint varies, it is important to see each feature's global impact, which is shown in Figure 6.

This SHAP value plot ranks the features in decreasing order of importance, while the spread across the horizontal determines the impact on the model for higher values (in red) and lower values (in blue). As is shown in Figure 6, the energy features are ranked high. Lower values of both binding energies are characteristic of the positive dataset. Hence, lower values of the binding energy tend to result in a positive impact on the model output; here, it is the classification in the positive dataset.

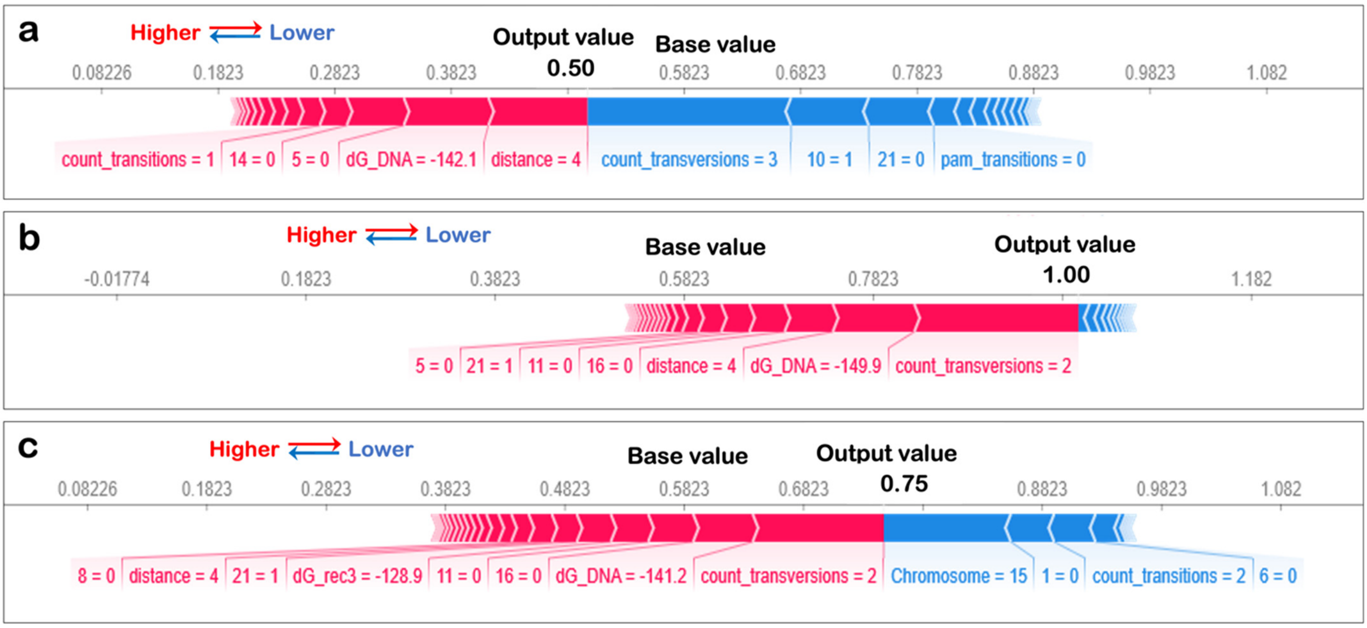

Figure 5. SHAP value plots for singular datapoints. Examples shown are for dataset indices (**a**) 10, (**b**) 17 and (**c**) 21, and are chosen randomly. The base value shown increases by features shown in red and decreases because of features shown in blue. Each feature impacts the value in magnitude indicated by SHAP values labelled alongside for each instance.

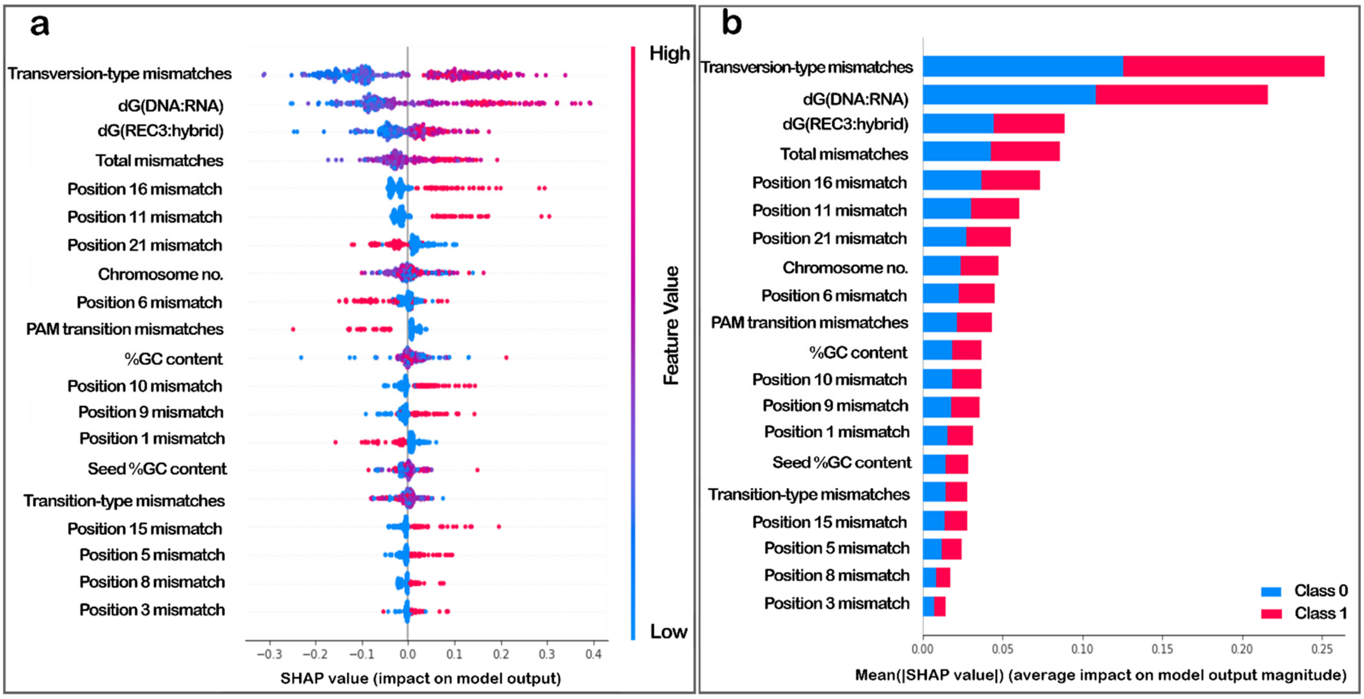

Figure 6. (**a**) SHAP value plot indicating global impact n model output. Each dot is an instance for a datapoint, the colour represents if the value for that instance is low (blue) or high (red). The spread indicates the magnitude of impact on the model output. (**b**) SHAP summary plot shows the impact of the features on each model output, negative class shown in blue and positive class shown in red, as stacked bars, in decreasing order of impact on output.

4. Discussion

The accurate prediction of CRISPR/Cas9 activity is crucial to not only designing experiments for various applications but also understanding the mechanism of Cas9 activity in vivo. Computational methods for predicting activity, off-targets and guide design have advanced significantly in recent times, yet there remains room for improvement regarding precision and accuracy. Prediction models would also benefit from improved and more sensitive Cas9 off-target detection methods to better distinguish between sequences likely to be acted upon by Cas9 (here, the positive dataset). This study reported that the incorporation of novel features allows for creating reliable prediction models. Moreover, the identification of novel features also sheds light on the factors influencing Cas9 activity in vivo.

The two major binding events responsible for Cas9 activity are (1) the binding of the Cas9 protein to the guide RNA, allowing DNA interrogation for complementary sequences, (2) followed by binding to the complementary sequence, which allows nuclease activation and a subsequent DNA double-stranded break [47]. Significantly accelerated by the availability of X-ray and cryo-EM structures, computational methods, such as QM/MM and molecular dynamics (MD), have elucidated the pre-catalytic and catalytic structures of Cas9 [48,49]. Enhanced MD simulations have shed light on the concerted mechanisms of HNH and RuvC domain activities [50–52]. The HNH domain via an Mg^{2+} ion cuts the target strand, while the RuvC domain houses two metal ions coordinated by conserved residues, which mediate a break in the non-target strand [52]. The varying tolerance of the mismatches across the guide-target heteroduplex has also been investigated [18,53,54]. The REC3 domain is known to interact with the guide RNA-target DNA complex, investigate the complementarity between the two, and tolerate mismatches [55,56]. Mismatches were seen to be tolerated towards the centre of the guide–target hybrid [53]. In contrast, mismatches towards the end of the hybrid induced an extended opening of the heteroduplex and leading to a conformational lock with the "L2" loop region [54]. Hence, the interactions of the guide RNA with target DNA and the heteroduplex with the REC3 domain of Cas9 protein have been shown to play a decisive role in nuclease activation, leading to Cas9 activity. The introduction of mismatches alters the interactions, leading to altered Cas9 activity. Understanding the factors that govern the RNA:DNA interactions is critical to elucidating biological function that it is involved in [57–60]. Hence, to quantify the interactions, DNA-RNA hybrid binding energy and Cas9-hybrid binding energy were estimated and analysed. The scores were then included as features alongside sequence features, and machine learning models were built for Cas9 activity prediction. Well-performing models were selected to analyse the importance of the new energy-based features, if any.

The random forest algorithm outperformed the others tested on both classification and regression tasks. The improved performance could be attributed to the limited number of features on each split. When compared to individual decision trees, which have a higher bias, random forests tend to perform better because of the variance reduction due to bagging. The features used, as the results describe, have minimum redundancy. The energy features prove vital in driving model output in both regression and classification tasks. This feature importance was also observed in the second-best performing classification model: a support vector-based machine classifier (a second regressor was not evaluated due to the performance being subpar, not reliable enough to study feature importance). The importance of the number of mismatches in the seed region has already been established in multiple studies [61,62]. Interestingly, a higher number of transversions was shown not to be tolerated in the experimental dataset, indicating a preference in the sequences (Figure 6a). However, a bigger dataset is required to be tested to establish this. The "distance" feature's trend may also be inferred intuitively since lower values of total mismatches are likely to be observed in the positive dataset. The energy features' contribution was novel and ranked high consistently in multiple results, enough to be considered important. The performance of the reported random forest classifier was also compared against existing methods for off-target prediction and was found to perform better (Figure S4).

5. Conclusions

In this study, the binding energy of the Cas9 REC3 domain and the 20-nucleotide DNA-RNA hybrid, and the binding energy of the 20 nucleotides of target DNA to guide RNA were novel features and proposed to be important for Cas9 activity. In the regression model, which predicts Cas9 cleavage frequency, and the classification model, which predicts Cas9 activity, both these features were shown to be important in driving model output. The same importance of the features was observed in the classification model, which can reliably distinguish between experimentally likely and unlikely off-target sequences. The other features used in the model were standard features used in most studies: the number and position of mismatches and type of mismatch, among others. The binding energy features were not redundant and did not show correlation with the other features, and hence they can be implemented in future algorithms for improved off-target prediction and guide-RNA design algorithms.

Supplementary Materials: The following supporting information [23,24,28,63–65] can be downloaded at: https://www.mdpi.com/article/10.3390/biom12081123/s1, Figure S1: Features of the predicted dataset. On the vertical axis is the number of predicted sequences and on the horizontal axis is the number of experimentally validated sequences. (a) The number of CRISPcut predictions vs. CIRCLE-seq off-targets which shows poor correlation as can be determined by the low R^2 value of 0.22. Each dot represents a unique target sequence for which the number of experimentally validated off-targets, plotted on the *X*-axis are compared against the number of predicted off-targets using the CRISPcut tool, as plotted on the *Y*-axis (b) Shows only the number of accessible predictions plotted against number of CIRCLE-seq off-target sites for a guide, high correlation denoted by an R^2 value of 0.84 can be observed here; Figure S2: Correlation plot of the features used for model training. The dark blue diagonal indicates self-correlation. There is a poor correlation between most feature pairs, but a few high correlation islands in dark blue and yellow colour can be seen. Since cell lines are mutually exclusive, the correlation between the cell lines will be negative. The dark blue islands are between PAM mismatches, PAM transitions and PAM mismatch positions, which can be expected; Figure S3: The mean absolute error (MAE) multiplied by 10, and R^2 value plotted for each model tested, various models were tested with increasing n_estimators and random states. The dashed grey line marks the maximum R2 and minimum error instance, which corresponds to n_estimators of 18 and a random state of 6; Figure S4: The best-performing random forest classifier was compared with the existing off-target prediction models for predicting the off-targets of a randomly selected EMX1 locus. The precision was calculated against experimentally validated sequences obtained from CIRCLE-seq. The off-targets were obtained from the CRISPOR [1,2], CRISTA [3], Elevation [4], ge-CRISPR [5] and CRISPcut [6] webservers (accessed on 13 June 2021); Table S1: Results of the two sample Mann–Whitney U test; Table S2: Random forest classification model performance summary; Table S3: Details of negative dataset; Table S4: Complete set of features used in the model learning process.

Author Contributions: D.S.V.: conceptualization, methodology, formal analysis, writing—original draft; Y.V.: investigation, software; D.S.: conceptualization, supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All input files and python scripts used for the data generation and analysis are available on GitHub (https://github.com/TeamSundar/crispr-cas9-dG-study).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012**, *337*, 816–821. [CrossRef]
2. Cong, L.; Ran, F.A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P.D.; Wu, X.; Jiang, W.; Marraffini, L.A.; et al. Multiplex genome engineering using crispr/cas systems. *Science* **2013**, *339*, 819. [CrossRef] [PubMed]
3. Mali, P.; Yang, L.; Esvelt, K.M.; Aach, J.; Guell, M.; DiCarlo, J.E.; Norville, J.E.; Church, G.M. RNA-guided human genome engineering via cas9. *Science* **2013**, *339*, 823–826. [CrossRef] [PubMed]
4. Porteus, M. Genome editing: A new approach to human therapeutics. *Annu. Rev. Pharmacol. Toxicol.* **2016**, *56*, 163–190. [CrossRef] [PubMed]
5. Gasiunas, G.; Barrangou, R.; Horvath, P.; Siksnys, V. Cas9–crrna ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 15539. [CrossRef] [PubMed]
6. Garneau, J.E.; Dupuis, M.-È.; Villion, M.; Romero, D.A.; Barrangou, R.; Boyaval, P.; Fremaux, C.; Horvath, P.; Magadán, A.H.; Moineau, S. The crispr/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **2010**, *468*, 67–71. [CrossRef]
7. Nishimasu, H.; Ran, F.A.; Hsu, P.D.; Konermann, S.; Shehata, S.I.; Dohmae, N.; Ishitani, R.; Zhang, F.; Nureki, O. Crystal structure of cas9 in complex with guide rna and target DNA. *Cell* **2014**, *156*, 935–949. [CrossRef]
8. Hsu, P.D.; Scott, D.A.; Weinstein, J.A.; Ran, F.A.; Konermann, S.; Agarwala, V.; Li, Y.; Fine, E.J.; Wu, X.; Shalem, O. DNA targeting specificity of rna-guided cas9 nucleases. *Nat. Biotechnol.* **2013**, *31*, 827. [CrossRef] [PubMed]
9. Fu, Y.; Foden, J.A.; Khayter, C.; Maeder, M.L.; Reyon, D.; Joung, J.K.; Sander, J.D. High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells. *Nat. Biotechnol.* **2013**, *31*, 822. [CrossRef] [PubMed]
10. Tsai, S.Q.; Zheng, Z.; Nguyen, N.T.; Liebers, M.; Topkar, V.V.; Thapar, V.; Wyvekens, N.; Khayter, C.; Iafrate, A.J.; Le, L.P. Guide-seq enables genome-wide profiling of off-target cleavage by crispr-cas nucleases. *Nat. Biotechnol.* **2015**, *33*, 187. [CrossRef]
11. Tsai, S.Q.; Nguyen, N.T.; Malagon-Lopez, J.; Topkar, V.V.; Aryee, M.J.; Joung, J.K. Circle-seq: A highly sensitive in vitro screen for genome-wide crispr–cas9 nuclease off-targets. *Nat. Methods* **2017**, *14*, 607. [CrossRef] [PubMed]
12. Wang, X.; Wang, Y.; Wu, X.; Wang, J.; Wang, Y.; Qiu, Z.; Chang, T.; Huang, H.; Lin, R.-J.; Yee, J.-K.J.N.b. Unbiased detection of off-target cleavage by crispr-cas9 and talens using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **2015**, *33*, 175–178. [CrossRef]
13. Wienert, B.; Wyman, S.K.; Richardson, C.D.; Yeh, C.D.; Akcakaya, P.; Porritt, M.J.; Morlock, M.; Vu, J.T.; Kazane, K.R.; Watry, H.L.J.S. Unbiased detection of crispr off-targets in vivo using discover-seq. *Science* **2019**, *364*, 286–289. [CrossRef] [PubMed]
14. Kim, D.; Kim, J.-S.J.G.r. Dig-seq: A genome-wide crispr off-target profiling method using chromatin DNA. *Genome Res.* **2018**, *28*, 1894–1900. [CrossRef] [PubMed]
15. May, A.P.; Cameron, P.; Settle, A.H.; Fuller, C.K.; Thompson, M.S.; Cigan, A.M.; Young, J.K. SITE-Seq: A Genome-Wide Method to Measure Cas9 Cleavage. 2017. Available online: https://protocolexchange.researchsquare.com/article/nprot-5889/v1 (accessed on 12 July 2022).
16. Doench, J.G.; Hartenian, E.; Graham, D.B.; Tothova, Z.; Hegde, M.; Smith, I.; Sullender, M.; Ebert, B.L.; Xavier, R.J.; Root, D.E. Rational design of highly active sgrnas for crispr-cas9–mediated gene inactivation. *Nat. Biotechnol.* **2014**, *32*, 1262. [CrossRef] [PubMed]
17. Doench, J.G.; Fusi, N.; Sullender, M.; Hegde, M.; Vaimberg, E.W.; Donovan, K.F.; Smith, I.; Tothova, Z.; Wilen, C.; Orchard, R. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nat. Biotechnol.* **2016**, *34*, 184. [CrossRef] [PubMed]
18. Klein, M.; Eslami-Mossallam, B.; Arroyo, D.G.; Depken, M.J.C.r. Hybridization kinetics explains crispr-cas off-targeting rules. *Cell Rep.* **2018**, *22*, 1413–1423. [CrossRef] [PubMed]
19. Xu, X.; Duan, D.; Chen, S.-J. Crispr-cas9 cleavage efficiency correlates strongly with target-sgrna folding stability: From physical mechanism to off-target assessment. *Sci. Rep.* **2017**, *7*, 143. [CrossRef]
20. Cui, Y.; Xu, J.; Cheng, M.; Liao, X.; Peng, S. Review of crispr/cas9 sgrna design tools. *Interdiscip. Sci. Comput. Life Sci.* **2018**, *10*, 455–465. [CrossRef] [PubMed]
21. Yennmalli, R.; Kalra, S.; Srivastava, P.A.; Garlapati, V.K. Computational tools and resources for crispr/cas 9 genome editing method. *MOJ Proteom. Bioinform.* **2017**, *5*, 00164.
22. Lin, J.; Wong, K.-C. Off-target predictions in crispr-cas9 gene editing using deep learning. *Bioinformatics* **2018**, *34*, i656–i663. [CrossRef]
23. Listgarten, J.; Weinstein, M.; Kleinstiver, B.P.; Sousa, A.A.; Joung, J.K.; Crawford, J.; Gao, K.; Hoang, L.; Elibol, M.; Doench, J.G. Prediction of off-target activities for the end-to-end design of crispr guide rnas. *Nat. Biomed. Eng.* **2018**, *2*, 38–47. [CrossRef]
24. Abadi, S.; Yan, W.X.; Amar, D.; Mayrose, I. A machine learning approach for predicting crispr-cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS Comp. Biol.* **2017**, *13*, e1005807. [CrossRef]
25. Wang, J.; Zhang, X.; Cheng, L.; Luo, Y. An overview and metanalysis of machine and deep learning-based crispr grna design tools. *RNA Biol.* **2020**, *17*, 13–22. [CrossRef]
26. Chuai, G.; Ma, H.; Yan, J.; Chen, M.; Hong, N.; Xue, D.; Zhou, C.; Zhu, C.; Chen, K.; Duan, B. Deepcrispr: Optimized crispr guide rna design by deep learning. *Genome Biol.* **2018**, *19*, 80. [CrossRef]
27. Luo, J.; Chen, W.; Xue, L.; Tang, B. Prediction of activity and specificity of crispr-cpf1 using convolutional deep learning neural networks. *BMC Bioinform.* **2019**, *20*, 332. [CrossRef] [PubMed]

28. Dhanjal, J.K.; Radhakrishnan, N.; Sundar, D. Crispcut: A novel tool for designing optimal sgrnas for crispr/cas9 based experiments in human cells. *Genomics* **2019**, *111*, 560–566. [CrossRef] [PubMed]
29. Liu, Q.; Di He, L.X. Prediction of off-target specificity and cell-specific fitness of crispr-cas system using attention boosted deep learning and network-based gene feature. *PLoS Comp. Biol.* **2019**, *15*, e1007480. [CrossRef] [PubMed]
30. Alkan, F.; Wenzel, A.; Anthon, C.; Havgaard, J.H.; Gorodkin, J. Crispr-cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biol.* **2018**, *19*, 177. [CrossRef]
31. Jensen, K.T.; Fløe, L.; Petersen, T.S.; Huang, J.; Xu, F.; Bolund, L.; Luo, Y.; Lin, L. Chromatin accessibility and guide sequence secondary structure affect crispr-cas9 gene editing efficiency. *FEBS Lett.* **2017**, *591*, 1892–1901. [CrossRef] [PubMed]
32. Chen, Y.; Zeng, S.; Hu, R.; Wang, X.; Huang, W.; Liu, J.; Wang, L.; Liu, G.; Cao, Y.; Zhang, Y. Using local chromatin structure to improve crispr/cas9 efficiency in zebrafish. *PLoS ONE* **2017**, *12*, e0182528. [CrossRef] [PubMed]
33. Uusi-Mäkelä, M.I.; Barker, H.R.; Bäuerlein, C.A.; Häkkinen, T.; Nykter, M.; Rämet, M. Chromatin accessibility is associated with crispr-cas9 efficiency in the zebrafish (danio rerio). *PLoS ONE* **2018**, *13*, e0196238. [CrossRef] [PubMed]
34. Jacobson, M.P.; Friesner, R.A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002**, *320*, 597–608. [CrossRef]
35. Sastry, G.M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221–234. [CrossRef]
36. DeLano, W.L. Pymol molecular viewer: Updates and refinements. In *Abstracts of Papers of the American Chemical Society*; American Chemical Society: Washington, DC, USA, 2009.
37. Jacobson, M.P.; Pincus, D.L.; Rapp, C.S.; Day, T.J.; Honig, B.; Shaw, D.E.; Friesner, R.A. A hierarchical approach to all-atom protein loop prediction. *Proteins Struct. Funct. Bioinform.* **2004**, *55*, 351–367. [CrossRef]
38. Genheden, S.; Ryde, U. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **2015**, *10*, 449–461. [CrossRef]
39. Kollman, P.A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; et al. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897. [CrossRef] [PubMed]
40. Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the mm/pbsa and mm/gbsa methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Modeling* **2011**, *51*, 69–82. [CrossRef] [PubMed]
41. Vallat, R. Pingouin: Statistics in python. *J. Open Source Softw.* **2018**, *3*, 1026. [CrossRef]
42. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
43. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 2017, Long Beach, CA, USA, 19 May 2017; pp. 4765–4774.
44. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
45. Dhanjal, J.K.; Dammalapati, S.; Pal, S.; Sundar, D. Evaluation of off-targets predicted by sgrna design tools. *Genomics* **2020**, *112*, 3609–3614. [CrossRef] [PubMed]
46. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.-W.; Newman, S.-F.; Kim, J.; et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760. [CrossRef] [PubMed]
47. Jiang, F.; Doudna, J.A. Crispr–cas9 structures and mechanisms. *Annu. Rev. Biophys.* **2017**, *46*, 505–529. [CrossRef]
48. Jiang, F.; Taylor, D.W.; Chen, J.S.; Kornfeld, J.E.; Zhou, K.; Thompson, A.J.; Nogales, E.; Doudna, J.A. Structures of a crispr-cas9 r-loop complex primed for DNA cleavage. *Science* **2016**, *351*, 867–871. [CrossRef]
49. Huai, C.; Li, G.; Yao, R.; Zhang, Y.; Cao, M.; Kong, L.; Jia, C.; Yuan, H.; Chen, H.; Lu, D. Structural insights into DNA cleavage activation of crispr-cas9 system. *Nat. Commun.* **2017**, *8*, 1375. [CrossRef]
50. Zhao, L.N.; Mondal, D.; Warshel, A. Exploring alternative catalytic mechanisms of the cas9 hnh domain. *Proteins Struct. Funct. Bioinform.* **2020**, *88*, 260–264. [CrossRef]
51. Casalino, L.; Nierzwicki, Ł.; Jinek, M.; Palermo, G. Catalytic mechanism of non-target DNA cleavage in crispr-cas9 revealed by ab initio molecular dynamics. *ACS Catal.* **2020**, *10*, 13596–13605. [CrossRef]
52. Palermo, G. Structure and dynamics of the crispr–cas9 catalytic complex. *J. Chem. Inf. Modeling* **2019**, *59*, 2394–2406. [CrossRef]
53. Mitchell, B.P.; Hsu, R.V.; Medrano, M.A.; Zewde, N.T.; Narkhede, Y.B.; Palermo, G.J.F.i.m.b. Spontaneous embedding of DNA mismatches within the rna: DNA hybrid of crispr-cas9. *Front. Mol. Biosci.* **2020**, *7*, 39. [CrossRef]
54. Ricci, C.G.; Chen, J.S.; Miao, Y.; Jinek, M.; Doudna, J.A.; McCammon, J.A.; Palermo, G.J.A.c.s. Deciphering off-target effects in crispr-cas9 through accelerated molecular dynamics. *ACS Cent. Sci.* **2019**, *5*, 651–662. [CrossRef] [PubMed]
55. Nierzwicki, Ł.; Arantes, P.R.; Saha, A.; Palermo, G. Establishing the allosteric mechanism in crispr-cas9. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*, e1503. [CrossRef] [PubMed]
56. Bravo, J.P.K.; Liu, M.-S.; Hibshman, G.N.; Dangerfield, T.L.; Jung, K.; McCool, R.S.; Johnson, K.A.; Taylor, D.W. Structural basis for mismatch surveillance by crispr–cas9. *Nature* **2022**, *603*, 343–347. [CrossRef] [PubMed]
57. Cheatham, T.E.; Kollman, P.A. Molecular dynamics simulations highlight the structural differences among DNA: DNA, rna: Rna, and DNA: Rna hybrid duplexes. *J. Am. Chem. Soc.* **1997**, *119*, 4805–4825. [CrossRef]

58. Nadel, J.; Athanasiadou, R.; Lemetre, C.; Wijetunga, N.A.; Broin, P.Ó.; Sato, H.; Zhang, Z.; Jeddeloh, J.; Montagna, C.; Golden, A. RNA: DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenet. Chromatin* **2015**, *8*, 46. [CrossRef]
59. Palermo, G. Dissecting structure and function of DNA rna hybrids. *Chem* **2019**, *5*, 1364–1366. [CrossRef]
60. Terrazas, M.; Genna, V.; Portella, G.; Villegas, N.; Sánchez, D.; Arnan, C.; Pulido-Quetglas, C.; Johnson, R.; Guigó, R.; Brun-Heath, I. The origins and the biological consequences of the pur/pyr DNA· rna asymmetry. *Chem* **2019**, *5*, 1619–1631. [CrossRef]
61. Semenova, E.; Jore, M.M.; Datsenko, K.A.; Semenova, A.; Westra, E.R.; Wanner, B.; Van Der Oost, J.; Brouns, S.J.; Severinov, K. Interference by clustered regularly interspaced short palindromic repeat (crispr) rna is governed by a seed sequence. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 10098–10103. [CrossRef]
62. Boyle, E.A.; Andreasson, J.O.; Chircus, L.M.; Sternberg, S.H.; Wu, M.J.; Guegler, C.K.; Doudna, J.A.; Greenleaf, W.J. High-throughput biochemical profiling reveals sequence determinants of dcas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 5461–5466. [CrossRef]
63. Haeussler, M.; Schönig, K.; Eckert, H.; Eschstruth, A.; Mianné, J.; Renaud, J.-B.; Schneider-Maunoury, S.; Shkumatava, A.; Teboul, L.; Kent, J.; et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **2016**, *17*, 148. [CrossRef]
64. Concordet, J.-P.; Haeussler, M. CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **2018**, *46*, W242–W245. [CrossRef] [PubMed]
65. Kaur, K.; Gupta, A.; Rajput, A.; Kumar, M. ge-CRISPR—An integrated pipeline for the prediction and analysis of sgRNAs genome editing efficiency for CRISPR/Cas system. *Sci. Rep.* **2016**, *6*, 30870. [CrossRef] [PubMed]