

**MECHANISTIC INSIGHTS INTO RNA-GUIDED
GENOME EDITING NUCLEASES**

DHVANI SANDIP VORA



**DEPARTMENT OF BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI**

MARCH 2023

© Indian Institute of Technology Delhi (IITD), New Delhi, 2023

Certificate

This is to certify that the thesis titled '**Mechanistic Insights into RNA-guided Genome Editing Nucleases**' being submitted by **Ms. Dhvani Sandip Vora** to the Indian Institute of Technology Delhi for the award of the degree of '**Doctor of Philosophy**', is a record of the bonafide research work carried out by her, which has been prepared under my supervision in conformity with the rules and regulations of the Indian Institute of Technology Delhi. The research reports and the results presented in this thesis have not been submitted for any degree or diploma in any other University or Institute.

Dr. D. Sundar

Institute Chair Professor

Department of Biochemical Engineering and Biotechnology

Indian Institute of Technology Delhi

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. D. Sundar, for his patient guidance and support by giving his time, ideas and feedback. I could only complete the journey of my PhD because of the motivation provided by him. I have improved by trying to imbibe his scientific and personal principles.

I am grateful to Prof. Ashok Kumar Patel, Prof. Ravikrishnan Elangovan and Prof. Preeti Srivastava, the members of my Students' Research Committee, for providing support and expertise in shaping this thesis. Completing this endeavor would not have been possible without Dr. Jaspreet Kaur Dhanjal, who has greatly influenced my professional and personal growth during my time as a PhD student.

I express my gratitude to UGC for the financial support during my PhD and to IIT Delhi for having a student-friendly and supportive administration. I also appreciate the IITD Kailash hostel staff and mess staff for their hospitality. Special thanks to my colleagues and friends from the lab- Dr. Vidhi Malik, Dr. Neetu Singh, Dr. Atul Jaiswal, Dr. Moolchand, Shashank Yadav, Yugesh Verma, Sakshi Bhandari, Navaneethan Radhakrishnan, Yogesh Kalakoti, Vipul Kumar, Seyad Shefrin and Pragya Kesarwani; I have learned a lot from each of them on both scientific and personal fronts. I have been able to broaden my perspective and find friendships in many other visiting scholars, undergraduate and graduate project students in my research group.

I also acknowledge the DAAD for the binational PhD fellowship and express gratitude to Jun.-Prof. Michael Boettcher for hosting me during my stay at the Martin Luther University, Germany. The training to perform CRISPR/Cas9 related experiments has deepened my understanding of the subject and broadened my perspectives.

I would like to acknowledge my parents and sister for their love and encouragement in this challenging pursuit; I could strive for betterment with the support of their faith and affection. I would also like to thank my friends, fellow PhD students and hostel mates who made my time at IIT Delhi an enjoyable experience.

Dhvani Sandip Vora

Abstract

Efficient manipulation of genes in organisms has propelled research in understanding mechanisms that govern life. New techniques that allow editing in single-celled organisms as well as complex life forms such as animals and plants are gaining traction. At the heart of such advances are custom-designed nucleases like zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and the CRISPR/Cas9 system.

A naturally occurring adaptive defence mechanism found in prokaryotes, the CRISPR/Cas9 system has been repurposed as an RNA-guided DNA targeting machinery. This transformative technology has shown great promise for biology, genetics and medicine. However, its widespread application for safe and effective genome editing and transcription modulation has been curbed by its inadequate specificity. In this thesis, the key factors involved in determining the efficacy of CRISPR/Cas9 have been studied that can be incorporated in guide RNA design tools to improve their predictions. Further, the study of the mechanism of on-target activity by Gaussian accelerated molecular dynamics solution has been carried out to determine sub-molecular interactions essential for driving Cas9 activity.

सार

जीवों में जीनों के कुशल क्रम परिवर्तन ने जीवन को नियंत्रित करने वाले तंत्रों को समझने में अनुसंधान को प्रेरित किया है। नई तकनीकें जो एकल-कोशिका वाले जीनोम के साथ-साथ जानवरों और पौधों जैसे जटिल जीवन रूपों में संपादन की अनुमति देती हैं, आकर्षण प्राप्त कर रही हैं। इस तरह की प्रगति के केंद्र में अनुकूलित किए गए न्यूक्लियोज़ जैसे ज़िंक फिंगर न्यूक्लियोज़ (TALEN), ट्रांसक्रिप्शन एक्टिवेटर-लाइक इफ़ेक्ट न्यूक्लीज़ (ZFN) और क्रिस्पर-कस 9 सिस्टम हैं।

प्रोकैरियोट्स में पाया जाने वाला एक स्वाभाविक रूप से होने वाला अनुकूली रक्षा तंत्र, क्रिस्पर-कस 9 प्रणाली को एक आरएनए-निर्देशित डीएनए लक्ष्यीकरण मशीनरी के रूप में पुनर्निर्मित किया गया है। इस परिवर्तनकारी तकनीक ने जीव विज्ञान, आनुवंशिकी और चिकित्सा के लिए बहुत आशाजनक कार्य किये हैं। हालाँकि, यह सुरक्षित और प्रभावी जीनोम संपादन और प्रतिलेखन समायोजन के लिए व्यापक अनुप्रयोग है, इसकी अपर्याप्त विशिष्टता प्रतिरोध बन गयी है।

प्रस्तुत शोध प्रबन्ध में क्रिस्पर-कस 9 की प्रभावशीलता का निर्धारण करने में शामिल प्रमुख कारकों का अध्ययन किया गया है। जिन्हें उनकी भविष्यवाणियों को बेहतर बनाने के लिए गाइड डिज़ाइन टूल में शामिल किया जा सकता है। इसके अलावा, कस 9 गतिविधि को चलाने के लिए आवश्यक उप-आणविक, परस्पर प्रभाव निर्धारित करने के लिए गॉसियन त्वरित आणविक गतिशीलता समाधान द्वारा ऑन-टारगेट गतिविधि के तंत्र का अध्ययन किया गया है।

Table of Contents

List of figures.....	xiii
List of tables.....	xxvii
List of abbreviations	xxx
1 Introduction to CRISPR/Cas9.....	1
1.1. Reprogrammable Nucleases	3
1.1.1. Zinc Finger Nucleases	3
1.1.2. Transcription activator-like effector nucleases	4
1.2. CRISPR/Cas9.....	5
1.3. Challenges in using CRISPR/Cas9 system as a tool for gene editing	7
1.4. Definition of Problem	9
1.5. Objectives	9
1.6. Thesis organization	10
2 Improving Guide RNA Design	14
2.1. To improve the sgRNA design tool for better off-target prediction for sgRNA.....	15
2.1.1. Background.....	15
2.1.2. Methods	16
(i) Data preprocessing.....	16
(ii) Design and training a multilayer perceptron.....	17
(iii) Model performance evaluation	18
2.1.3. Results.....	19
2.1.4. Conclusions.....	21
2.2. Identifying position-wise nucleotide preference and tolerance to mismatches in sgRNA	22
2.2.1. Background.....	22
2.2.2. Methods	24
(i) Data assembly- positive off-target.....	24
(ii) Data assembly- negative off-target.....	24

(iii)	sgRNA-target sequence encoding.....	25
(iv)	Mismatch and bulge propensity analysis	25
(v)	Architecture of the recurrent convolutional neural network (CRISP-RCNN).....	26
(vi)	Model Training	27
(vii)	Evaluation metrics	27
(viii)	Identification of important sequence features.....	28
2.2.3.	Results.....	30
(i)	Data assembly and preparation	30
(ii)	Analysis of mismatch occurrence and nucleotide-specific mismatch propensity.....	30
(iii)	Analysis of bulge occurrence and nucleotide-specific bulge propensity.....	35
(iv)	Predictive performance of the CRISP-RCNN model	37
(v)	Activation Maps (AMs)	39
(vi)	Average AMs feature importance	40
2.2.4.	Conclusions.....	42
3	Machine Learning-based off-target prediction and evaluation using novel features	45
3.1.	Background.....	45
3.2.	Importance of energy-based features	45
3.2.1.	Objectives	45
3.2.2.	Methods	46
(i)	Selecting sequences for positive and negative datasets	46
(ii)	System preparation and binding energy calculation	46
(iii)	Predictive features.....	47
(iv)	Mann-Whitney U statistic	47
(v)	Machine learning model implementation	48

(vi) Sampling data for training	49
(vii) Assessing model performance	49
(viii) Assessing feature importance	50
3.2.3. Results.....	50
(i) Binding energy-based features.....	50
(ii) Mann-Whitney U test.....	52
(iii) Regression model and SHAP analysis.....	54
(iv) Classification model and SHAP analysis.....	58
3.2.4. Conclusions.....	62
3.3. Importance of DNA shape and cell line features	62
3.3.1. Background.....	62
3.3.2. Methods	63
(i) Compiling positive and negative datasets.....	63
(ii) Sequence featurization	64
(iii) DNA shape feature determination	64
(iv) Cell line feature extraction.....	64
(v) Model training.....	65
(vi) Performance evaluation	66
(vii) Model selection.....	67
(viii) Determination of feature importance	68
3.3.3. Results.....	69
(i) Data compilation and featurization.....	69
(ii) Model performance	69
(iii) Model selection.....	71
(iv) Feature importance analysis.....	73
3.3.4. Conclusions.....	76
4 Molecular Dynamics of Cas9 with insights into mechanism of activation and catalysis	78

4.1. Background.....	78
4.2. Methods	79
4.2.1. Structure preparation.....	79
4.2.2. Molecular dynamics simulations	81
4.2.3. Essential dynamics analysis.....	84
4.2.4. Conformational entropy	85
4.2.5. Nucleic acid hybrid stability	85
4.2.6. Linear interaction energy	86
4.3. Results.....	86
4.3.1. Conformational dynamics analysis	90
4.3.2. Essential dynamics analysis.....	94
4.3.3. Effects of Cas9 mutants on gRNA-tDNA stability.....	96
4.3.4. Critical interactions between Cas9 variants and the gRNA-tDNA hybrid.....	101
4.4. Conclusions.....	107
5 Conclusions and Future Prospects	109
References.....	112
Appendix.....	123
Publications from the thesis	131
Resume of the author.....	133

List of Figures

Figure 1: Double-stranded break (DSB) repair outcomes. Repair can either be mediated by either non-homologous end joining (NHEJ), in the absence of a template, or if a template strand is present, homology directed repair (HDR). Image adapted from Joung and Sander, (2013) [20].....3

Figure 2: Zinc Finger Nucleases. Arrays of zinc finger domains fused to *FokI* nuclease domain can recognize specific sequences of DNA and create a double-stranded break.4

Figure 3: Transcription Activator-Like Effector (TALE) domains, identifying one DNA base each, arranged in an array (shown in blue colour) and covalently linked to the nuclease domain of *FokI* to recognize and cleave a particular stretch of DNA.....5

Figure 4: Mechanism of Cas9-sgRNA target recognition and cleavage. The 20-nucleotide guide RNA enables Cas9 to identify target DNA to be cleaved, resulting in blunt-end cuts 3 bases downstream to the PAM site.7

Figure 5: Off-target activity demonstrated by Cas9. The sgRNA, shown in green, binds not only to target site with perfect complementarity, shown in blue, but also to unintended sites with a mismatch and with deletions, shown in orange.8

Figure 6:One-Hot encoding. A schematic representation of how one-hot encoding works in the case of DNA. A 4-row matrix is made, where 1 indicates presence of a nucleotide, while zeros represent absence. In one column, there will be a single one and three zeros, the positions of which are decided by the nucleotide present at that position.17

Figure 7: Class imbalance in the dataset. The zero (0) represents negative off-target set, derived from *CRISPCut* [63], while the one (1) represents positive off-target set, derived from GUIDE-Seq and CIRCLE-Seq [76, 77]. 17

Figure 8: Cross-validation. Schematic diagram of the k-fold cross-validation strategy. One portion of the data is fed to the model for training, one portion is used for validation during training. An unseen portion of the data is then used to test the model. 18

Figure 9: Confusion matrix. The columns correspond to the true class and the rows represent the predicted class. The diagonal cells are correctly classified samples, the others are incorrect classifications. The cells in white are total samples of a class, both correctly and incorrectly classified. The bottom right grey cells show overall accuracy. The confusion matrices shown are for the training data (a), validation data (b), test data (c) and all data (d). 20

Figure 10: Receiver Operating Characteristics (ROC) curves for the training data (a), validation data (b), test data (c) and all data (d). The ROC plots true positive rate (sensitivity) against false positive rate (1-specificity), since all the curves are close to the upper-left corner, the model can be said to be performing well. 21

Figure 11: Sequence to image encoding scheme. The 4 channels correspond to the 4 nucleotides. Like the one-hot encoding scheme, the presence of a nucleotide marks a positive in that channel, while the other channels show no values. Here, a monochromatic scheme has been used. N-padding has been used to equalize the off-target sequence lengths. At all positions of ‘N’, which may be any nucleotide, a 25% intensity fill is used for all 4 channels. 25

Figure 12: CRISP-RCNN model architecture. Representative target and off-target sequences are depicted as input for the model. The input is fed into convolution layers, followed by an LSTM layer. The two branches are then concatenated and fed into two

separate by fully connected dense layers, which perform the classification and regression tasks.26

Figure 13: Trend in mismatch tolerance across the length of the protospacer. (a)

Mismatch tolerance measured on the experimental dataset. (b) Mismatch tolerance weighted by experimental cleavage frequency. (c) Occurrence of insertions or deletions, represented as gaps, across the length of the protospacer.31

Figure 14: Nucleotide-specific trend in mismatch tolerance - (a) counted for all

instances when the target nucleotide is ‘A’, and (b) frequency-weighted trend for mismatch tolerance when ‘A’ is the target nucleotide, calculated individually for each position.32

Figure 15: Gap tolerance across the length of the guide RNA. The fraction of gaps

at each position among all positive off-target sequences is shown in the figure, weighted by the frequency of occurrence in the experiment. The stacked column plot indicates the percentage of positions at which mismatches are found in the dataset. As can be observed, the gaps are solely clustered towards the centre of the 20-nucleotide sequence while the mismatches are towards the flanks.36

Figure 16: Predictive performance of various neural networks tested. Three

architectures were tested for performing simultaneous classification and regression tasks on the dataset- convolutional neural networks (CNN) shown in violet, CNN with an LSTM layer (CNN-LSTM), shown in pink, and CNN with a bidirectional LSTM layer (CNN-biLSTM), shown in green. The three models were optimized, the best performing models were selected and evaluated on independently for 5 runs. The average scores of the area under Precision-Recall curve (auPR), minor class recall, Matthew’s correlation coefficient (MCC) was plotted to compare the classification performance and the R^2 and minor class R^2 were compared for the regression

performance. The non-parametric Kruskal-Wallis test was performed to determine if the differences among the scores were significant (* indicates $p < 0.01$). The classification performance was comparable, except for the auPR scores, where the CNN-biLSTM outperformed the other architectures. The regression performance of the CNN-biLSTM was better than the other two architectures, hence, was selected for further analyses.37

Figure 17: CRISP-RCNN prediction performance. (a) The precision-recall (PR) curve with an area under the curve (AUC) of 0.89, (b) the F1-MCC curve indicating robust performance even on the imbalanced dataset, the blue points indicating best- and worst-possible prediction performance. In plots (a) and (b) the dashed line indicates random prediction, i.e., 50% accuracy; and (c) the confusion matrix indicating the distribution of predicted against the true labels.38

Figure 18: Activation maps (AMs). (a) The FANCF locus target and a representative positive off-target sequences are considered. The sequence is numbered from the PAM-distal end towards the ‘NGG’ PAM site. The N-padding is numbered in negative. (b) DeepSHAP, (c) gradient explainer (GradExp) and (d) smoothed saliency maps are depicted. The regions marked in red indicate positive contribution and regions marked in blue indicate a negative contribution for (b) and (c). In the case of smoothed saliency (d), only regions that contribute to the assignment to the class are marked in red, regardless of positive or negative contribution.39

Figure 19: Average activation maps. EMX1 locus has been shown as a representative example. (a) The EMX1 target sequence. Average activation maps of the (b) positive off-target sequences, (c) positive class reference target sequences, (d) negative off-target sequences, and (e) negative class reference target sequences.41

Figure 20: Predicted dataset features. The number of predicted sequences is plotted on the vertical axis and the number of experimentally validated sequences is plotted on the horizontal axis. (a) is the number of *CRISPCut* predictions vs. CIRCLE-seq off-targets which shows poor correlation as can be determined by the low R2 value of 0.22. (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 R2 value of 0.84 can be observed here.51

Figure 21: Correlation plot of the features used for model training. The dark blue diagonal indicates self-correlation. There is poor correlation between most feature pairs but a few high correlation islands in dark blue and yellow colour can be seen. Since the cell lines are mutually exclusive, correlation between the cell lines will be negative. The other dark blue islands are between PAM mismatches, PAM transitions and PAM mismatch positions, which can be expected.52

Figure 22: The Mean Absolute Error (MAE) multiplied by 10 and R2 values are 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 6.55

Figure 23: SHAP variable importance plots. (a) The plot arranges features in a 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 color indicates feature value for that output.56

Figure 24: SHAP variable importance plots for a singular datapoint. 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 is influenced by varying degrees by the features shown in the diagrams and the output value (shown in bold) is 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.....57

Figure 25: 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) and (b). In (c) vertical clusters at individual values indicate correlation with dG(DNA:RNA) values and the plot also shows negative correlation of the values of the distance with the output variable.....58

Figure 26: Performance metrics of the classifier. (a) Confusion Matrix for the random forest classifier. The 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, which has an area under the curve of 0.94 for the whole dataset (c) receiver operating characteristic (ROC) for the test dataset, which plots the true positive rate vs. false positive rate, the area under the curve (AUC) is 0.95. The dashed line shows 50% accuracy.....59

Figure 27: SHAP value plots for a singular datapoint. Examples shown are for dataset indices (a) 10, (b) 17 and (c) 21, have been 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.60

Figure 28: SHAP plots for the classifier. (a) SHAP value plot indicating global impact on 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 magnitude of impact on the model output. (b) SHAP summary plot shows the features impact 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.61

Figure 29: Performance of neural networks trained on three datasets. The nets were trained on a varying number of trainable model parameters, plotted on the horizontal axis. The three datasets used for training are the Sequence (in blue), Sequence and Epigenetic (in orange), and Sequence, Epigenetic and Shape (in gray) with the same output variables. The various measures of model performance compared are (a) model loss, (b) area under the precision-recall curve (auPR), (c) precision, and (d) recall, all measured on the test datasets.....70

Figure 30: Model performance comparison with previous reported algorithms. (a) The various metrics of accuracy, precision, recall, specificity, negative predictive value (NPV) and area under the receiver operating characteristic curve (AUC_ROC) are compared for the selected classifiers from the presented study (nn1, nn57 and nn79) with previously reported algorithms- CRISTA, OffTargetPred (OffTPred) and CFD scoring [113-115]. (b) Receiver operating characteristic (ROC) plots comparing the models selected with the previously reported off-target prediction methods. The dashed line indicating random prediction.72

Figure 31: Feature importance analysis. (a) LIME feature importance ranking. The bar in the X-axis direction indicates the value of the feature importance as calculated

using the LIME method. Top 10 features are shown for reference. POI6 indicates insertion at position 6. OTA3 stands for on-target base “A” (adenine) at position 3. (b) Consensus feature importance. The permutation importance ranks for the train dataset against the test dataset. The bubble size indicates the LIME feature importance. The size legend indicates bubble sizes for LIME rank 1 and LIME rank 10 for reference. The importances are plotted for the three classifiers selected- nn1 (yellow), nn57 (blue) and nn79 (pink). The prominent bubbles are labelled with the feature numbers, overlapping bubbles of the same feature are labelled only once. 74

Figure 32: The three Cas9 variants investigated. The target DNA (tDNA) shown in blue is identified by the guide RNA (gRNA) depicted in pink bound to the three variants (a) wtCas9, (b) eSpCas9 and (c) hypaCas9 proteins in grey. The amino acid mutations introduced in the wildtype that confer the improved efficiency are demonstrated as a ball-like green representation for both variants- eSp and hypaCas9. 80

Figure 33: Comparison of parmbsc0 and parmbsc1 forcefields. The figure (a) shows residue-wise fluctuation while figure (b) shows time-dependent fluctuation of the molecule for both parmbsc0, indicated in blue, and parmbsc1, indicated in red. 87

Figure 34: (a) H-bond plot. The number of hydrogen bonds made between the protein and the nucleic acids per frame have been plotted. **(b) RMSD plot.** The root-mean square deviation per frame of the protein backbone has been plotted. 88

Figure 35: The PMF (potential of mean force) of the S867-N1054 (left) and S355-S867 (right) FRET distances. The plots describe the conformational change of the HNH domain from the RNA-bound to the catalytically active state. FRET distances derived from experimental values are shown by vertical lines. Figure adapted from [129]. 88

Figure 36: Inter-residue distance and time-averaged distance during GaMD simulations of wt Cas9. Distance between N1054- S867 (a) and S867- S355 (b) are plotted against time of simulation. Time averaged distances between the residues N1054- S867 (c) and S867- S355 (d) are expected to converge but no such convergence is observed in either plot.89

Figure 37: RMSD plots of the Cas9-nucleic acid complexes. (a) SpCas9-DNA-sgRNA complex and (b) DNA-sgRNA hybrids of *wt* Cas9 shown in green, eSpCas9 shown in orange and HypaCas9 shown in purple.90

Figure 38: RMSD plots of the various components of the simulated system. (a) Stabilized protein RMSD calculated from a single trajectory depicting convergence of the system. (b) The gRNA-tDNA hybrid, based on the RMSD profile, when bound to hypaCas9 is observed to show marginally more fluctuations than when bound to the wtCas9 and eSpCas9. The differences in the various Cas9-bound hybrid RMSDs can be attributed to the PAM-distal region (c), the PAM-proximal region shows comparable RMSD for the three systems (d).92

Figure 39: Root mean square fluctuation (RMSF) plots. The RMSF of the guide RNA (gRNA)- target DNA (tDNA) hybrids of the three variants- wt Cas9 (blue), HypaCas9 (red) and eSpCas9 (green). Residues numbered 1-10 are those belonging to the PAM proximal ends of the guide (g) RNA and target (t) DNA. Residues numbered 10-20 are the PAM-distal nucleotides of the RNA-DNA hybrid.93

Figure 40: Principal component analysis of the three Cas9 variants- wt, eSp and hypa. (a) The first three principal components are plotted and coloured with a gradient depicting a frame-wise progression. (b) The first and second principal components are plotted for the wtCas9 (green), eSpCas9 (orange) and hypaCas9 (purple) (c) the second and third components are plotted. The differences among the Cas9 mutants based on

the dynamic spread is better understood in the projections, indicating differences in the structural ensembles sampled in one trajectory for each variant.95

Figure 41: Principal component analysis of the three Cas9 variants and the bound gRNA-tDNA hybrids.

(a) The projections of the first three principal components are coloured according to a gradient depicting a frame-wise progression in one trajectory.

(b) Porjection of the first two principal components for wtCas9 (green), eSpCas9 (orange) and hypaCas9 (purple), and similarly (c) the projection of the second and third principal components were analysed. In contrast to the protein principal components projections, the heteroduplex projections show stark variations in conformational spread among the three variant Cas9-bound hybrid trajectories.....97

Figure 42: Base pair parameters calculated and the trends observed.

(a) The structural distortion determined from different base pairing parameters is represented schematically. The distortion measured across the simulation time in base pairing of the gRNA-tDNA bound to wtCas9 (green), eSpCas9 (orange) and hypaCas9 (purple) are measured in terms of the opening (b) and stretch (c). The central line is the median, the box boundaries represent the 25th and 75th percentile. The 75th percentile of the data, the 3rd quartile are the outlier boundaries. All other outliers, exceeding the outlier whiskers are shown as distinct diamond points. In case the fluctuations are too large to be plotted, the distinct points are the outliers and the median and quartiles are not shown.98

Figure 43: Base pair parameters calculated across the first trajectory for the hybrids bound to the three variants, across the length of the hybrid.

The parameters included are (a) buckle, (b) propeller, (c) shear and (d) stagger. The letter-value plot indicates the distribution of the data around a median (indicated by a grey line across the box), with the outliers plotted as diamonds. The trend observed in all the parameters

calculated demonstrates a stability across the hybrid, except the three PAM-distal bases, though complementary. The distortions were fewer in the eSpCas9-bound hybrid and the most distortions were observed in the hypaCas9-bound hybrid.99

Figure 44: The hybrid base pairing distortions measured for last four of the PAM-distal bases across the simulation time of one trajectory. The distortions are measured in terms of the (a) buckle, (b) opening, (c) propeller, (d) shear, (e) stagger and (f) stretch. The parameters for the three Cas9 variants are plotted- wtCas9 (green), hypaCas9 (violet) and eSpCas9 (orange). The flat line at zero indicates that the distortions are too large to be calculated..... 100

Figure 45: Probability density functions of interaction energies. (a) The electrostatic component of the total interaction energy is compared for the three variant systems based on the probability density function (pdf) plotted for wtCas9 (green), eSpCas9 (orange) and hypaCas9 (purple). The pdf plot shape and peak shows substantial differences in the protein-hybrid electrostatic interaction energy for the three variant systems. (b) The interaction energy of the Cas9 protein and PAM-proximal segments of the tDNA-gRNA hybrid shows negligible differences in the pdf plot. (c) Contrarily, the energy of interaction of the Cas9 protein and PAM-distal regions show variations..... 105

Figure 46: Analysis of the interaction energy based on the van der Waals component by comparing the probability density functions of the three Cas9 variant-bound hybrids. Little variation is observed for the pdf of the three Cas9 variants- wt (green), eSp (orange) and hypa (purple). The PAM-proximal contributions (b), and PAM-distal contributions (c) in the hybrid are equivalent, with the exception of the broadening of the peak observed for the PAM-distal region of the hypaCas9-

hybrid, indicating small variations in the structures sampled during the simulation time.

..... 106

Appendix Figure 1: Mismatch tolerance across the length of the guide RNA. The fraction of sites that remain unsubstituted are illustrated, i.e., they do not tolerate mismatches in the positive off-targets evaluated. The plots also illustrate the type of substitutions tolerated in each case when the target nucleotides are ‘C’ (a) and (b), ‘G’ (c) and (d), and ‘T’ (e) and (f). The figures (a), (c) and (e) are for when the number of mismatches are calculated from the sequences once per occurrence. The plots (b), (d) and (f) are for the fraction of mismatches are calculated weighted on the frequency of occurrence in the experiment. 123

Appendix Figure 2: Position-wise trends in gaps and mismatches. The trend is depicted for the target nucleotides (a) adenine, “A” shown in green, (b) cytosine, “C” shown in yellow, (c) guanine, “G” shown in blue and (d) thymine “T” shown in pink. The line indicating percentage gaps observed at each position is plotted in dark red.

..... 124

Appendix Figure 3: LIME feature importance for the top 50 features. The features ranked 11 through 50 are depicted in this chart, in continuation of the main Figure 3(a), which lists the top ten ranked features. The importance values depict impact of the features on model output..... 125

Appendix Figure 4: Schematic representation of the intra-base pair parameters. (a) Stretch, (b) Opening, (c) Propeller, (d) Shear, (e) Stagger and (f) Buckle; and inter-base pair parameters (g) Shift, (h) Slide, (i) Rise, (j) Tilt, (k) Roll and (l) Helix Twist (HelT). Features used for the study but not included in the image are Major groove width (MGW) and electrostatic potential (EP). 126

Appendix Figure 5: Model performance comparison on including novel features.

The performance of similar model architectures is compared on inclusion of three different features of the same dataset- sequence (S), epigenetic (E) and DNA shape (D) features. The classifier performance is also measured in three other metrics- (a) and (d) Precision, (b) and (e) area under the Precision-Recall curve (auPR), and (c) and (f) Kappa score. All the scores reported are measured on the unseen test dataset. The plots (a), (b) and (c) depict model performance on S+E+D features and S+E features when plotted against scores for only S features. The dashed line represents equivalent performance. The plots (d), (e) and (f) show the trend in predictive power with change in number of model training parameters. The increase in performance on the datasets with DNA shape is evident from the plots..... 127

Appendix Figure 6: Domain-wise RMSD analysis of the three variants of Cas9 under study.

The RMSD is calculated across one simulation trajectory for the (a) recognition, i.e., REC domains, (b) the PAM-interacting C-terminal domain, (c) the HNH nuclease domain and (d) RuvC nuclease domain..... 128

Appendix Figure 7: The distortions in the hybrid base-pair parameters measured for a PAM-proximal base.

The geometric parameters were measured for the second base in the hybrid, across the simulation time of one trajectory, as a reference for parameters indicating stable base pairing. The plots indicate the deviation from the expected (plotted along the zero on the vertical axis), and all values lie along the expected with minor deviation. These plots provide a reference of the deviation that can be accounted for..... 129

List of Tables

Table 1: Summary of model architecture and accuracies. Multiple models were tested, and the accuracies obtained have been reported. The default feedforward neural network was then optimized in terms of nodes of the hidden layer (12) and regularization was introduced.	19
Table 2: Position-nucleotide pairs with least mismatch tolerance. Column1 lists the position at which the expected target nucleotide is retained the most, i.e., sequences with no mismatches and for the same nucleotide-position pair, when the expected nucleotide is not retained, the most likely found nucleotide at that position are mentioned.....	33
Table 3: Base-position pairs that tolerate the most mismatches. For each position at which a nucleotide was found to be retained least is summarized in the “Position” and “Target nucleotide” columns. For each position-nucleotide pair when the expected target nucleotide is the least favoured, the most commonly found nucleotide is also mentioned.....	34
Table 4: Gap tolerance across the length of guide RNA. The positions at which notable number of gaps were observed (>0.005% of all match and mismatch occurrences).	35
Table 5: Performance scores for various metrics of the best-performing model (CRISP-RCNN), measured on the held-out test dataset.	38
Table 6: Details of negative dataset preparation. The negative dataset was assembled with reference to the 8 unique guides selected from the CIRCLE-seq dataset [77]. ...	46

Table 7: Results of the two sample Mann-Whitney U test. The H_0 hypothesis is that the two groups have comparable values, rejected H_0 hypothesis indicates the differences in random values selected from the two groups is statistically significant. P-value is less than 0.01 indicating less error and hence, confidence in the test. The Observed standard effect size is the difference in probability to choose a bigger value from the the negative dataset and positive dataset. The probability that a random value selected from the negative set is larger than one selected randomly from the positive set is denoted by common language effect size.....53

Table 8: Summary of model performances. All values shown are on the test dataset.54

Table 9: Classification models performance summary. The accuracy values mentioned are for the test set.59

Table 10: Model performance of the random forest classifier. Measured on test dataset. The accuracy reported is after 5-fold cross validation.....59

Table 11: Experiment-wise breakdown of various experimentally verified off-target data sources. The number of cell lines and the number of unique guides used in the individual experiment is mentioned. The italicised numbers indicate that the experiment was carried out with genomic DNA *in vitro*.64

Table 12: Experimental values of inter-residue distances of different states of Cas9 [129]......89

Table 13: Conformational entropies calculated for the first (S1) and second (S2) trajectories. The trend of conformational entropy calculated remains consistent in both trajectories. The wild-type (wt) Cas9 demonstrating the highest entropy. The conformational entropies of the variants are comparable, indicating fewer accessible conformations.94

Table 14: Loss of interaction between the Cas9 recognition lobe and the gRNA-tDNA hybrid. The base labels for gRNA bases begin with R and those for tDNA begin with D..... 101

Table 15: Loss of interaction between Cas9 RuvC domain and the gRNA-tDNA hybrid. A significant loss of Hydrogen bonds between the Cas9 RuvC domain and hybrid at the PAM distal end was observed, when comparing the variants eSpCas9 and HypaCas9 with the wtCas9..... 102

Table 16: Loss of interactions observed between Cas9 HNH domain and the gRNA-tDNA hybrid. Loss of multiple H-bonds has been observed between the hybrid and the Cas9 HNH nuclease domain of HypaCas9 when compared to eSpCas9 and wtCas9..... 103

Table 17: Loss of interaction between Cas9 L1 linker and the gRNA-tDNA hybrid. No loss of interactions between L1 and hybrid in eSpCas9, HypaCas9 as compared to wtCas9 were observed. However, significant gain in interactions were observed in eSpCas9 and HypaCas9. The new interactions observed in eSpCas9 are between ASN767 – RU10 (83.76 %), GLN771 – RU9 (57.53 %). The additional interactions observed for HypaCas9 are between ARG765 – DT13 (85.05 %), ARG765 – DG14 (92.37 %), THR769 – RU10 (67.89 %). Additional investigation would be required to determine the role these interactions..... 103

Appendix Table 1: Complete set of features used in the model learning process.
..... 130

List of Abbreviations

AMs	Activation Maps
AUC	Area Under the Curve
Cas9	CRISPR-associated nuclease 9
CRISPR	Clustered Regularly Interspersed Palindromic Repeats
crRNA	CRISPR RNA
DSB	Double-stranded break
GaMD	Gaussian accelerated molecular dynamics
HDR	Homology-directed repair
MLP	Multilayer Perceptron
NHEJ	Non-homologous end joining
nt	nucleotide
PAM	Protospacer Adjacent Motif
pdf	Probability Density Function
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
sgRNA	single guide RNA
SHAP	SHapley Additive exPlanations
TALENs	Transcription Activator-like Nucleases
tDNA	target DNA
tracrRNA	trans-CRISPR RNA
ZFNs	Zinc Finger Nucleases