Safety quantification in gene editing experiments using machine learning on rationally designed feature spaces



Florian M. Störtz
Linacre College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary Term 2023

Th' infinite Creator can dwell in it,
and may not man?
—Joseph Beaumont (1616-1699), *The House of the Mind*

Contents

Abstract iv

Acknowledgements vi

List of publications arising from this work viii

List of Figures ix

List of Tables xiv

List of Abbreviations xvi

1 Background **1**

1.1 Biological background 1

1.1.1 Clinical gene therapy 1

1.1.2 CRISPR 3

1.1.3 Cleavage repair 8

1.1.4 DNA structure & energetics 9

1.2 Computational background 11

1.2.1 Gradient boosted decision trees 12

1.2.2 Neural Networks 13

1.2.3 Convolutional Neural Network 14

1.2.4 Recurrent Neural Network 15

1.2.5 MOSAICS 18

1.3 Existing CRISPR modelling efforts 18

1.3.1 Cleavage data repositories 18

1.3.2 sgRNA specificity modelling efforts . . . 18
1.3.3 Existing editing outcome prediction algorithms . . . 21
1.3.4 Industry review . . . 22
1.3.5 Conclusion . . . 23

2 Database platform for CRISPR/Cas9 off-target cleavage assays 24

2.1 Introduction . . . 24
2.2 Materials and Methods . . . 26
2.2.1 Data Acquisition . . . 26
2.2.2 Data storage . . . 32
2.2.3 Website interface . . . 32
2.3 Results and Discussion . . . 34
2.3.1 Future Developments . . . 37

3 Physically informed modelling of CRISPR/Cas9 cleavage efficiency 40

3.1 Introduction . . . 41
3.2 Exploratory modelling . . . 41
3.2.1 Data Source . . . 42
3.2.2 Data Augmentation . . . 46
3.2.3 Feature Encoding . . . 48
3.2.4 Model Architectures . . . 49
3.2.5 Model Training & Evaluation . . . 51
3.2.6 Results & Discussion . . . 56
3.2.7 Conclusion . . . 63
3.3 Optimised selection of architectures, feature sets and encodings . . . 68
3.3.1 Model Training & Evaluation . . . 69
3.3.2 Results & Discussion . . . 70
3.3.3 Optimal split of data into train & test set . . . 81

4 CRISPR repair outcome prediction using structural calculations **86**

4.1 Introduction 86

4.2 Preliminary considerations 87

4.2.1 Insights from the FORECasT dataset 88

4.2.2 Approach 1: Adjust residue twist angle according to indel length . . . 89

4.2.3 Approach 2: Sweeping indel position across the cut site 92

4.2.4 Conclusion 93

4.3 CasTER repair steps modelling 94

4.3.1 Thermodynamic model of DSB repair 94

4.3.2 Data source 95

4.3.3 Computational data generation 96

4.3.4 Feature engineering 98

4.3.5 Modelling 98

4.3.6 Results & Discussion 101

5 Conclusion & Further Developments **111**

Bibliography **114**

A Work on other publications **125**

A.1 Analysis of epigenetic descriptors 125

B Physically informed modelling: additional data **141**

C Repair outcome modelling: additional data **149**

Abstract

With ongoing development of the CRISPR/Cas programmable nuclease system, applications in the area of *in vivo* therapeutic gene editing are increasingly within reach. However, non-negligible off-target effects remain a major concern for clinical applications. Even though a multitude of off-target cleavage datasets have been published, a comprehensive, transparent overview tool has not yet been established. The first part of this thesis presents the creation of crisprSQL (http://www.crisprsql.com), a large, diverse, interactive and bioinformatically enhanced collection of CRISPR/Cas9 off-target cleavage studies aimed at enriching the fields of cleavage profiling, gene editing safety analysis and transcriptomics.

Having established this data source, we use it to train novel deep learning algorithms and explore feature encodings for off-target prediction, systematically sampling the resulting model space in order to find optimal models and inform future modelling efforts. We lay emphasis on physically informed features which capture the biological environment of the cleavage site, hence terming our approach piCRISPR. We find that our best-performing model highlights the importance of sequence context and chromatin accessibility for cleavage prediction and compares favourably with state-of-the-art prediction performance. We further show that our novel, environmentally sensitive features are crucial to accurate prediction on sequence-identical locus pairs, making them highly relevant for clinical guide design.

We then turn our attention to the cell-intrinsic repair mechanisms that follow CRISPR/Cas-induced cleavage and provide a prediction algorithm for the outcome genotype distribution based on thermodynamic features of the DNA repair process. In a pioneering approach, we utilise structural calculations for the generation of these features and show that this novel approach surpasses published outcome prediction algorithms within our testing regime. Through interpretation of the trained model, we elucidate the thermodynamic factors driving DNA repair and provide a computational tool that allows experts to assess the severity

of the genotypic changes predicted for a given edit.

Together, these efforts provide a comprehensive, one-stop computational source to assess and improve CRISPR/Cas9 gene editing safety.

Structure of this Thesis

The objective of this research work is to increase understanding of key biological and physical mechanisms involved in CRISPR/Cas cleavage and subsequent double strand break repair by means of data, bioinformatics and computational modelling.

Chapter 1 outlines the relevant background concerning clinical gene therapy using CRISPR/Cas, and provides an introduction into the machine learning methods used in this work.

Chapter 2 then concerns the assembly and bioinformatic annotation of our cleavage training and testing dataset.

Chapter 3 describes our implementation of state-of-the-art cleavage prediction algorithms and how we use the dataset from Chapter 2 to train and evaluate these.

Chapter 4 presents our approach to utilising structural calculations to predict the outcomes of cell-intrinsic repair pathways which are triggered by CRISPR/Cas-induced cleavage. It therefore concludes the safety consideration of the CRISPR/Cas gene editing process.

Acknowledgements

For this work to have come to fruition, I am indebted to the support of a whole range of people, to whom this thesis is dedicated.

First and foremost, I want to thank my fellow group members Furkan Özden and Jeffrey K. Mak for fruitful collaboration in the later stage of this project. The Interdisciplinary Bioscience Doctoral Training Partnership programme for believing in me ahead of this project, for giving a framework and backbone to my research, as well as for providing a network of like-minded, highly acclaimed, personable and enthusiastic doctoral students. M for support and inspiration in the early stages of this project, and without whom none of this would have happened. The Choir of Trinity College, Cambridge, especially SDL, HP, JD, DM, for being a musical haven, for purpose and focus in stressful times, and for extending my vocabulary, musical and emotional. RE, RS and RC for their warm-hearted encouragement and expertise, and for helping me grow as an artist and a person. MH for steadfast, heartfelt, unflinching friendship. NM, EM and MB for sparking and nourishing my interest in mathematics, physics and computer science. My brother, my grandparents and, most of all, my parents for their unwavering love and support, and for being the bedrock on which all of this is built. No home from home would be possible without the one they provide.

Above all, my grateful thanks go to my supervisor Dr Peter Minary, whose enthusiastic, dedicated, personable and encouraging style of mentorship I have greatly benefitted from and admired from start to finish. His humble disposition, his far-reaching expertise, his optimism and his talent to turn around what I had thought of as ill success have been truly inspirational.

Dass dieses Projekt so zur Vollendung gekommen ist, verdanke ich der Unterstützung einer ganzen Reihe an Personen, denen ich die vorliegende Arbeit widmen möchte.

Zuerst danke ich meinen Kollegen Furkan Özden und Jeffrey K. Mak für fruchtbare Zusam-

menarbeit im fortgeschrittenen Teil dieses Projekts. Dem Interdisciplinary Bioscience Doctoral Training Partnership – Programm für das Vertrauen im Vorfeld dieses Projekts, für den interdisziplinären Rahmen meines Forschungsprojekts und für das Netzwerk aus gleichgesinnten, außerordentlich fähigen, sympathischen und enthusiastischen Mitdoktoranden. M für Unterstützung und Inspiration am Anfang dieses Projekts, ohne die ich es vermutlich nicht angetreten hätte. Dem Choir of Trinity College, Cambridge, besonders SDL, HP, JD, DM, für eine musikalische Heimat, für Muße und Fokus in stressigen Zeiten, für das musikalische und emotionale Vokabular, das ich dort kennenlernen durfte. RE, RS und RC für ihre warmherzige Ermutigung und Expertise, die mich als Musiker und als Mensch haben wachsen lassen. MH für treue, tief empfundene, unbeirrbare Freundschaft. NM, EM und MB, die mein Interesse an Mathematik, Physik und Informatik geweckt und gefördert haben. Meinem Bruder, meinen Großeltern und besonders meinen Eltern für Liebe und Unterstützung. Sie sind der Fels auf dem diese Arbeit steht, und keine zweite Heimat wäre denkbar ohne meine erste, die ich ihnen verdanke.

Mein größter Dank aber gilt meinem Betreuer und Mentor Dr. Peter Minary, dessen enthusiastischer, engagierter, persönlicher und ermutigender Stil mich von Anfang bis Ende dieser Arbeit sehr geprägt hat und weiterhin mit großem Respekt erfüllt. Seine Bescheidenheit, seine weit reichende Expertise, sein Optimismus und sein Talent, scheinbare Misserfolge in Motivation umzuwandeln, haben mich sehr inspiriert.

List of publications arising from this work

- Florian Störtz, Peter Minary: crisprSQL: a novel database platform for CRISPR/Cas off-target cleavage assays, *Nucleic Acids Research* **49**, 855 (2021)
- Jeffrey Mak, Florian Störtz, Peter Minary: Comprehensive computational analysis of epigenetic descriptors affecting CRISPR-Cas9 off-target activity, *BMC Genomics* **23**, 805 (2022)
- Florian Störtz, Jeffrey Mak, Peter Minary: piCRISPR: Physically Informed Deep Learning Models for CRISPR/Cas9 Off-Target Cleavage Prediction, *Artificial Intelligence in the Life Sciences* **3**, 100075 (2023)
- Florian Störtz, Peter Minary: CasTER: Prediction of Cas9-induced DSB repair using thermodynamic features (in preparation)

Publications not discussed in this thesis:

- Jeffrey Mak, Artemi Bendandi, José Augusto Salim, Ivan Mazoni, Fabio Moraes, Luiz Borro, Florian Störtz, Walter Rocchia, Goran Neshich, Peter Minary: Accurate prediction of CRISPR-Cas9 off-target activity by learning to utilize protein 3D nanoenvironment descriptors (submitted)

List of Figures

1.1 Interventional clinical trials started since 2019 which use CRISPR genome editing, as of February 2022. Adapted from [7]. 3

1.2 CRISPR as a part of the viral defense system of prokaryotes. 5

1.3 Visualisation of the Cas9–DNA binding, schematic (a) and using X-ray diffraction (b). 6

1.4 Means of template-free DSB repair. Images reproduced from [40]. 10

1.5 DNA double helix structure, with phosphate backbone in orange and base pairs coloured by atom type. Nucleobases are shown with individual atoms in a stick representation. 10

1.6 Schematic of a fully connected neural network with input vector $\mathbf{x} = (x_1, x_2, x_3)$, two hidden layers with 4 neurons each, and the scalar output $\hat{y}$. 14

1.7 Recurrent neural network architecture. 15

1.8 Recurrent neural network architecture from Figure 1.7. 16

1.9 Gated Recurrent Unit as characterised by equations 1.12 - 1.15, showing the update gate of vector $z^{(t)}$ and reset gate of vector $r^{(t)}$. 17

1.10 Timeline of web-based tools for sgRNA design (selection). Until 2019, this is a reproduction of Figure 2 from [61]. 19

1.11 Output of the inDelphi web service [77] for a certain example sequence. . . . 22

2.1 Overview of the crisprSQL database. 33

2.2 Overview of the website showing the included studies, together with metadata such as assay type, the total number of guides and guide-target pairs observed as well as the predominantly cleaved gene names. 38
2.3 Web interface of the database. 38
2.4 Nonlinear transformation of the cleavage rate distribution. 39

3.1 Biological environment of the cut site, including epigenetic markers. 45
3.2 Overview of piCRISPR feature encodings. 50
3.3 Target-mismatch encoding (E1) as described in Figure 3.2. 51
3.4 Validation benchmark values for various combinations of feature sets, encodings and models. 53
3.5 Performance comparison between piCRISPR models when tested on held out studies [90, 96, 97] whilst ommitting certain sets of features. 56
3.6 Training performance of the six models when gradually introducing training data study by study. 57
3.7 2D distance network representation of pairwise training performance. 57
3.8 Violin density plots of the StrongWeak BDM feature. 59
3.9 Comparison of piCRISPR models, using various methods, features and encodings. 60
3.10 Receiver operating (left) and precision-recall (right) curves underlying Figure 3.9. 61
3.11 Comparison between piCRISPR (model CNN_S5E2) and CRISPR-Net [67]. . 62
3.12 Base-pair resolved global SHAP values for the CNN_S5E2 classification model. 65
3.13 Distributions of mean squared error between cleavage activity and prediction score for the CRISPR-Net algorithm and our CNN_S5E2 model. 67
3.14 Distributions underlying Table 3.2. 67
3.15 Overview of our two chosen feature encodings and models. 69

3.16 Comparison of piCRISPR models with published algorithms. 73

3.17 Comparison between piCRISPR, CRISPR-Net [67] and CRISPR-IP [68]. 74

3.18 Distributions of the relative pairwise difference as defined in Table 3.3 for the DeepCRISPR algorithm (blue bars) and our 6×23 CNN model (orange bars). The distributions for other sets of features can be found in Figure B.8. 75

3.19 Global SHAP values for the 16×23 CNN classification model. 78

3.20 Base-pair resolved SHAP values for the 16×23 (left panel, see Figure B.6) and 6×23 (right panel, see Figure B.5) CNN classification models. 79

3.21 SHAP dependence plot for the $E_{\text{CRISPRoff}}$ feature. 80

3.22 Visualisation of the Hamming distance matrix between the clusters identified by the DBSCAN clustering [144] of guides based on their sequence, cell line, DSB quantification assay and delivery method. 83

3.23 Performance benchmarks of the 16×23 RNN classification model in testing scenario 1 (top two plots) and quantities characterising a train/test split (bottom two panels). 85

4.1 The measured editing outcome frequency for the first 50,000 outcomes across several thousand oligos in dependence of their indel length. 89

4.2 Measured indel rate plotted against the calculated potential energy difference between unstrained DNA and DNA which has been torsionally strained according to indel length. 90

4.3 Distribution of all the calculated ΔE_{bppt}, grouped according to the residue twist angle (bppt) templates used to calculate the respective values. 91

4.4 Sequences involved in calculating the loss of base-stacking interactions during the repair of a cut at the indicated site, resulting in the deletion of sequence Y. In Figure 4.5, we sweep the position of sequence Y relative to its measured value (black arrows) without changing the length of Y. 92

4.5 Resulting internal energy differences ΔE during outcome repair, sweeping the tentative indel position around the measured outcome (dark point) for two example indel data points. 92

4.6 Histograms for the ΔE percentile in which the measured indel location falls. 93

4.7 Repair step sequences according to the model of non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ) given in [152], here at the example of a 9 bp deletion from the FORECasT dataset [76]. These give rise to our CasTER outcome prediction model. 95

4.8 Schematic for our CasTER Cas9 DSB repair outcome prediction model. 100

4.9 Scatter plot of the true vs. predicted activity label of a given indel signature within our chosen portion of the FORECasT data set [76]. 103

4.10 Distributions of symmetrised Kullback-Leibler (KL) divergence between predicted and true outcome distribution. 104

4.11 SHAP values of the $E^{(2)}_{\mathrm{MMEJ}}$ features by length of the annealing microhomology. 105

4.12 Histogram of CasTER repair outcome frequency predictions by predicted deletion length. 106

4.13 SHAP feature importance values for the energy + sequence model. 107

4.14 SHAP dependence plot of all the $E^{(2)}_{\mathrm{MMEJ}}$ features combined. 109

4.15 Sequence logo (top row) and heatmap (bottom row) representations of the global SHAP feature importance value for the target sequence. This is shown for the sequence + energy model (**A**) and the sequence-only model (**B**). . . . 110

B.1 Spearman r data underlying Figure 3.7. Data has been gained using the CNN_S4E0 model. 141

B.2 SHAP values for the CNN_S5E2 model (Figure 3.12). 142

B.3 Base-pair resolved global SHAP values for the RNN_S4E3 classification model. Training and test set are identical to Figure 3.12. 144

B.4 SHAP values for the RNN_S4E3 classification model (Figure B.3). 145

B.5 Base-pair resolved global SHAP values for the 6×23 CNN classification model. Training and test set are identical to Figure 3.19. 146

B.6 Base-pair resolved global SHAP values for the 16×23 CNN classification model. Training and test set are identical to Figure 3.19. 147

B.7 Base-pair resolved global SHAP values for the 16×23 RNN classification model. Training and test set are identical to Figure 3.19. 148

B.8 Distributions underlying Table 3.3. 148

List of Tables

1.1 Selection of established DSB mapping technologies as presented in [33]. Technologies take place *in vitro* unless given otherwise. 7

1.2 Non-comprehensive overview of online databases related to CRISPR. 19

1.3 Computational methods for prediction of sgRNA specificity. Publications until 2020 are given as a reproduction of Table 2 from [61]. 20

1.4 Overview of CRISPR editing outcome predictors published to date. 21

2.1 Data sources included in crisprSQL (version 10 July 2020), chronologically ordered by publication date. 27

2.2 Availability of epigenetic markers by cell line in the ENCODE, SCREEN [107] and GEO [116] databases. 31

2.3 Arguments for the crisprspec.get_eng() function with which CRISPRspec energy values have been retrieved. 32

3.1 Abbreviations for feature sets S0-S5 used in this section. 52

3.2 Quantities gained on the subset of duplicate guide-target sequence pairs (testing scenario 3) using our piCRISPR CNN_S5E2 regression model and the CRISPR-Net model. 66

4.1 Visualisation of the 230 possible indel signatures within our chosen ranges for indel length and start position, both measured in base pairs. Each is modelled as one individual label in our multilabel regression model. 96

4.2 Specifications of IT systems that were used to run the MOSAICS thermodynamics calculations, as well as the observed runtimes. 97

4.3 Benchmark values when using our model CasTER and the FORECasT model [76] on our chosen portion of the FORECasT data set. 102

List of Abbreviations

ARC	Oxford University Advanced Research Computing
AU	area under
AWS	Amazon Web Services (commercial cloud computing)
bp	(nucleotide) base pairs
bppt	base pairs per turn of DNA
CA	CRISPR cleavage activity: fraction of CRISPR-induced DSBs (read counts) at the respective location over the total number of measured DSBs
CNN	Convolutional Neural Network
DSB	DNA double strand break
dsDNA	double stranded DNA
GRU	gated recurrent unit
KL	Kullback–Leibler divergence
LSTM	Long Short Term Memory Recurrent Neural Network
mh	DNA sequence microhomology
MMEJ	microhomology-mediated end joining
NHEJ	non-homologous end joining
nt	nucleotide
PAM	protospacer adjacent motif
PDB	protein data bank file format
PRC	Precision-recall curve
RNN	Recurrent Neural Network
RNP	ribonucleoprotein complex consisting of Cas9 protein and sgRNA
ROC	Receiver operating curve (true positive rate in dependence of false positive rate)
sgRNA	(single) guide RNA
ssDNA	single stranded DNA

Chapter 1

Background

1.1 Biological background

1.1.1 Clinical gene therapy

As any medical intervention, the efficacy and safety, i.e. potential toxicity, of gene editing therapy must be weighed up against potential adverse effects, as well as the risk profile of any individual patient. Clinical applications of genome editing differ in four key features [1]: the type of cells or tissues which are modified (such as somatic cells or heritable changes in germ cells), whether the delivery happens *in vivo* or *ex vivo*, whether the edit is done to treat or prevent a disease or to introduce new traits, and whether the edit concerns a single mutation or the expression/introduction of a whole gene function.

For the purpose of this work, we concentrate on nuclease-mediated gene editing via CRISPR/Cas9, which is described in more detail below. For this class of gene editing, a system consisting of a nuclease and targeting sequences has to be delivered inside cells. Common techniques for this are integration by viral vectors which allow permanent integration, physical methods such as electroporation and chemical methods such as lipofection or cell-penetrating peptides [2]. These all come with their own toxicity and immunogenicity profiles [3].

The mechanisms for such toxicity, leading to disadvantageous outcomes of CRISPR/Cas-induced mutagenesis, are elucidated by Davies [4] at the example of human germline editing in a 1-cell embryo. Besides mosaicism (a non-uniform distribution of mutations across the resulting organism), unwanted large deletions of several kilobases [5] and on-site damage of both mutant and normal alleles due to lack of sequence discrimination are named there. In reference [6], the authors show *in vitro* that these effects occur with considerable frequency. Furthermore, the immune response to the editing machinery could introduce further undesired effects, as could cellular apoptosis triggered by editing-induced double strand breaks [7].

Davies further refers to off-target effects at unwanted but sequence-similar loci on the genome [4] as drivers of editing-induced mutagenesis. Consequently, Dai *et al.* state that "one of the major hurdles to the clinical translation of CRISPR/Cas9 is its off-target effects" [8]. According to Fellman *et al.*, "off-target analyses [...] are far more critical for therapeutic CRISPR/Cas gene editing than for preclinical investigations" [9]. Sun *et al.* describe off-target effects as a "major obstacle for further applications of CRISPR/Cas9" and see the identification of all possible off-target sites as a "necessary step" for therapeutic application [10]. Researchers have directly observed leukaemia induction through genome editing with site-specific nucleases in human cells [11], as well as lesions distal of the cut site [5], highlighting the need for increased understanding of off-target effects.

Several clinical applications of gene editing technologies have already been realised to date, such as personalised cancer treatment with genetically modified chimeric antigen receptor (CAR) T cell-based therapies [9]. A variety of clinical trials to treat diseases using CRISPR genome editing-based interventions are currently ongoing (see Figure 1.1).

In this work, we concentrate on three issues causing CRISPR/Cas-induced mutagenesis named above, namely on-target efficiency, off-target specificity and post-edit cellular repair of double strand breaks. Evidence suggests that these phenomena can be explained to a

NCT number	Disease	Target gene and effect	Intervention	Study phase	(Estimated) study start	Status	Countries
Inherited blood disorders							
NCT04417764	Advanced HCC	*PD-1* knockout	Modified T cells	Phase I	Jun. 2019	Recruiting	China
NCT04035434	Refractory B-cell malignancies	*TCR* and *MHC-I* knockout, CAR insertion into *TRAC* locus	CD19-CAR-modified T cells lacking TCR and MHC-I	Phase I	Jul. 2019	Recruiting	United States, Australia, Canada, Germany, Spain
NCT04037566	Refractory B-cell malignancies	Disruption of HPK1	CD19-CAR-modified T cells with CAR delivered by lentivirus and Cas9 knockout of *HPK1*	Phase I	Aug. 2019	Recruiting	China
NCT04244656	Relapsed or refractory multiple myeloma	*TCR* and *MHC-I* knockout, CAR insertion into *TRAC* locus	BCMA-targeted CAR-T cell lacking TCR and MHC-I	Phase I	Jan. 2020	Recruiting	United States, Australia, Canada, Spain
NCT04426669	Metastatic gastrointestinal epithelial cancer	Cytokine-induced SH2 protein (CISH) knockout	Modified T cells	Phase I, Phase II	May 2020	Recruiting	United States
NCT04438083	Relapsed or refractory renal cell carcinoma	*TCR* and *MHC-I* knockout, CAR insertion into *TRAC* locus	CD70-CAR-modified T cells lacking TCR and MHC-I	Phase I	Jun. 2020	Recruiting	United States, Australia, Canada, Netherlands
NCT04502446	Relapsed or refractory T- or B-cell malignancies	*TCR* and *MHC-I* knockout, CAR insertion into *TRAC* locus	CD70-CAR-modified T cells lacking TCR and MHC-I	Phase I	Jul. 2020	Recruiting	United States, Australia, Canada
NCT04557436	B-cell leukemia	Modification of *CD52* and *TRAC*	CD19-CAR-modified T cells with CAR delivered by lentiviral vector and CRISPR-Cas9 editing of *CD52* and *TRAC* loci	Phase I	Aug. 2020	Recruiting	UK
NCT04767308	Relapsed/refractory CD5+ hematopoietic malignancies	*CD5* knockout	CD5-CAR T cells	Early Phase I	Mar. 2021	Not yet recruiting	—
NCT04637763	Relapsed/refractory B-cell NHL	*PD-1* knockout, CAR insertion into *TRAC* locus	CD19-CAR modified T cells	Phase I	May 2021	Recruiting	United States
NCT03690011	T-cell malignancies	*CD7* knockout in CD7 CAR-T cells	CAR-T cells to CD7 and knockout of native *CD7* to prevent self-targeting	Phase I	Aug. 2021	Recruiting	United States
NCT05066165	Acute myeloid leukemia	*TCRβ* and *TCRα* knockout, integration of WT1-TCR into *TRAC* locus	WT1-TCR modified T cells	Phase I, Phase II	Nov. 2021	Recruiting	United States, UK
NCT04976218	eGFR + solid tumors	*TGF-β receptor II* knockout	Modified T cells	Phase I	Dec. 2021	Not yet recruiting	—
NCT05037669	Relapsed or refractory B-cell malignancies	*B2M*, *CIITA*, and *TRAC* disruption	CD19-CAR-modified T cells lacking expression of HLA class I, HLA class II molecules, and endogenous	Phase I	Jan. 2022	Not yet recruiting	—
Metabolic disorders							
NCT04601051	ATTR amyloidosis	Disruption of *TTR*	*In vivo* gene disruption	Phase I	Nov. 2020	Recruiting	New Zealand, Sweden, UK
NCT05120830	Hereditary angioedema	*KLKB1* knockout	*In vivo* gene disruption	Phase I, Phase II	Nov. 2021	Recruiting	New Zealand
NCT05210530	Type 1 diabetes	CRISPR-Cas9-modified allogeneic pancreatic endoderm cells	Allogeneic *β*-cell replacement therapy derived from a CRISPR-Cas9 gene-edited immune-evasive stem cell line	Phase I	Jan. 2022	Recruiting	Canada
Others							
NCT03872479	Leber congenital amaurosis 10	Removal of alternative splice site in *CEP290*	*In vivo* gene correction	Phase I, Phase II	Sep. 2019	Recruiting	United States
NCT04560790	Refractory herpetic viral keratitis	HPV genome	Degradation of viral genome	Phase I, Phase II	Nov. 2020	Active, not recruiting	China
NCT04990557	COVID-19 respiratory infection	*PD-1* and *ACE2* knockout	Modified T cells	Phase I, Phase II	Aug. 2021	Not yet recruiting	—
NCT05144386	HIV infection	HIV genome	*In vivo* Removal of viral DNA	Phase I	Jan. 2022	Recruiting	United States
NCT05143307	HIV infection	HIV genome	*In vivo* Removal of viral DNA	Phase I	Mar. 2023	Enrolling by invitation	United States

Figure 1.1: Interventional clinical trials started since 2019 which use CRISPR genome editing, as of February 2022. Adapted from [7].

certain extent by sequence context and epigenetic markers, even though a full explanation would involve structural calculations about the heteroduplex molecules involved in the process [12, 13]. True structural insight is lacking for all tasks so far, which is an area in which we want to contribute expertise and software tools.

1.1.2 CRISPR

The clustered regularly interspaced short palindromic repeats (CRISPR) sequence family was first described in *E. coli* in 1987 [14], but it took until 2007 to recognise it as an adaptive

part of the viral defense system of most archaea ($\sim 90\%$) and bacteria ($\sim 40\%$) [15]. The defensive pathway has three steps [16] (see Fig. 1.2).

1. *Acquitision*: In the aforementioned prokaryotic organisms, invading exogenous viral DNA is firstly cleaved off by specialised nuclease enzymes. New spacers are derived from these so-called protospacers and integrated into the repeat-spacer array locus on the host genome. This is often located adjacently to the CRISPR locus and hence named CRISPR-associated (Cas). The CRISPR sequences therefore act as a viral history of the respective cell, and are stabilised by the palindromic nature of their saved states which results in stable secondary structures [17].

2. *Expression*: From there they can be transcribed to CRISPR RNAs (crRNAs) which, in type II CRISPR systems such as CRISPR/Cas9, hybridises with a trans-activating crRNA (tracrRNA).

3. *Interference*: The Cas endonuclease recognises invading copies of known viral DNA with the help of the crRNA and can subsequently render them inactive.

Researchers have repurposed this ability for programmable genome editing in many (also eukaryotic) species, complementing existing strategies such as zinc-finger nucleases (ZFNs, [18]) and transcription activator-like effector nucleases (TALENs, [19]).

Nuclease process For the purpose of this work, we concentrate on the effects of the Cas9 protein gained from *Streptococcus pyogenes*. The crRNA forms an active complex of about 50 bp length with the tracrRNA. Homology of the crRNA part with a 20 bp region in the genome results in annealing of a single-stranded overhang of the complex with this region. We can make use of this process by providing a *single guide RNA* (sgRNA) which mimics the crRNA-tracrRNA complex. We call the genomic region to which this binds the *target strand* (c.f. Figure 1.3). Binding is initiated by a favourable interaction of the 3 base pair long protospacer-adjacent motif (PAM) on the opposite, non-target strand with the Cas9 protein [16]. For *S. pyogenes* Cas9, this is the case for an 'NGG' PAM where N stands for an arbitrary

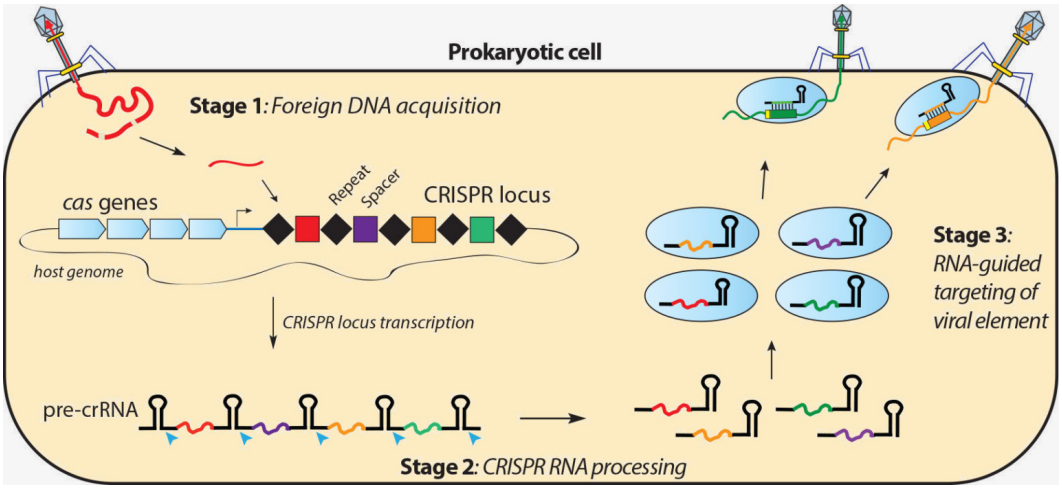
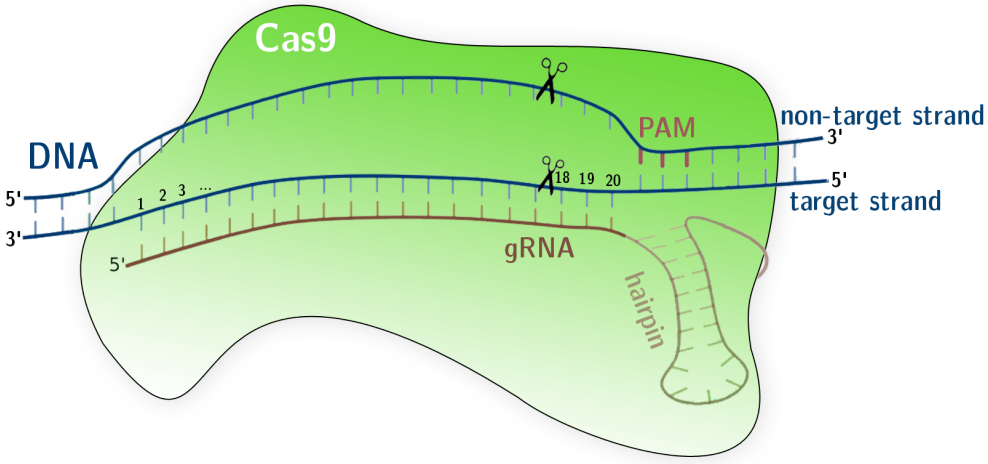


Figure 1.2: CRISPR as a part of the viral defense system of prokaryotes. Image courtesy of J. Doudna [20].

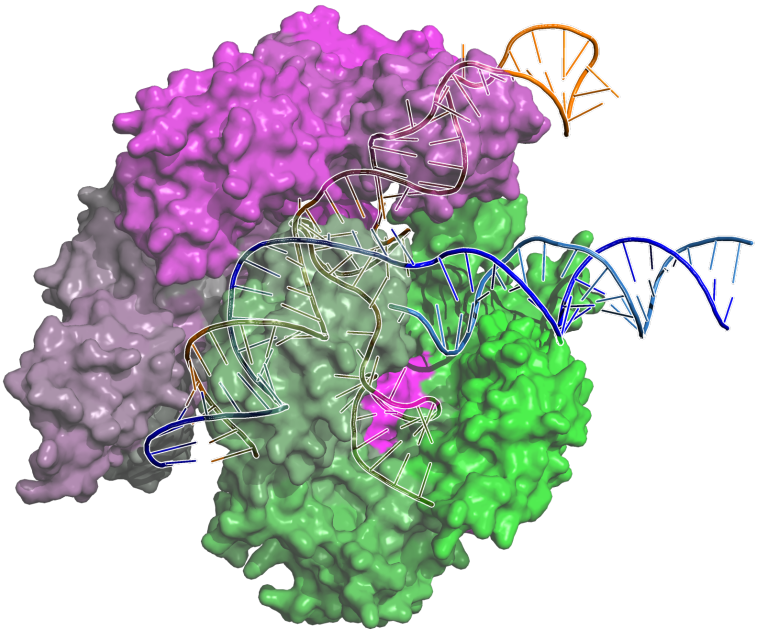
base (A, T, C, G). Binding at other, noncanonical PAMs such as 'NAG' and 'NGA' have also been observed [21]. A depiction of this can be found in Figure 1.3.

After binding has taken place, the nuclease domains of Cas9 (RuvC, HNH) can cleave the double-stranded DNA 3bp upstream of the PAM, leaving a *double-strand break* (DSB). Due to the stochastic, energy-driven nature of both the binding and the cleavage process, we expect a distribution of cuts over the whole genome, including undesired *off-target effects* which could possibly have catastrophic consequences as described above, e.g. by knocking out tumor suppressor genes like p53 and Rb [25]. It is therefore necessary to carefully design sgRNAs for use with a specific Cas nuclease, optimising on-target activity (efficiency) and off-target specificity. Both have to be considered for clinical applications and are influenced by the method of delivery, as well as the *ex vivo* or *in vivo* nature of the gene-editing step [7].

Targeting specificity The main driver of specificity are base-pairing interactions between the sgRNA and the target DNA strand. A successful base-pairing induces a conformational change in the HNH domain of the Cas nuclease (see Figure 1.3b) which is necessary for cleavage activity [16]. DNA strand separation (unwinding) and sgRNA–DNA heteroduplex formation start at the PAM-proximal "seed" region, which suggests that the base pair mismatch tolerance of the CRISPR process increases towards the PAM-distal region.



(**a**) The gRNA (brown) anneals with a single strand of the double-stranded DNA (blue), entering a so-called R-loop configuration. The non-annealing parts of the crRNA form a hairpin configuration. The Cas9 endonuclease enzyme (green) cleaves the double-stranded DNA at the indicated position, 3bp upstream of the protospacer adjacent motif (PAM). Schematic depiction based on [13].



(**b**) Crystal structure of Cas9 bound to a DNA target (blue) with matching PAM, from X-ray diffraction [22] (PDB: 4UN3). Coloured by functional groups, with the colours matching the ones in (a). The pink regions on the top are the HNH and RuvC domains cleaving the target and non-target strand, respectively [23, 24]. The bright pink arginine-rich region in the centre is responsible for PAM recognition [24].

Figure 1.3: Visualisation of the Cas9–DNA binding, schematic (a) and using X-ray diffraction (b).

CRISPR cleavage quantification

In order to obtain a comprehensive picture of the effect of a certain combination of sgRNA, CRISPR/Cas9 and cell line, a genome-wide DSB mapping technology is needed. A selective

overview of these can be found in Table 1.1.

name	working principle	notes
whole genome sequencing [26]	next-generation sequencing	unbiased, but detects only high-frequency mutations
GUIDE-seq [27], 2015	detects nuclease activity based on NHEJ-mediated integration of double-stranded oligodeoxynucleotides (dsODNs) at DSB sites	cell-based but not applicable *in vivo* due to the integration of exogenous DNA
Digenome-seq [21], 2015	count number of sequence reads with the same 5′ end	robust, established, relatively expensive
BLESS [28], 2016	direct labelling of breaks, enriched with high-throughput sequencing and streptavidin beads	cell-based, in situ labelling avoids detection of DSBs formed during DNA extraction
SITE-Seq [29], 2017	selective tagging & enrichment followed by sequencing	
CIRCLE-seq [30], 2017	intramolecular ligation circularises DNA, subsequently linearised by Cas9, increasing resolution for PCR	works without a reference genome
CHANGE-seq [31], 2020	Tn5 transposase enzyme tags DNA fragments, creates overhang to circularise DNA	faster and easier than CIRCLE-seq, data currently only available on primary T cells
Extru-seq [32], 2023	RNPs introduced into live cells, cells are killed to stop repair machinery	works without reference genome, no re-ligation of DSBs due to absence of NHEJ

Table 1.1: Selection of established DSB mapping technologies as presented in [33]. Technologies take place *in vitro* unless given otherwise.

The selection of existing DSB mapping technologies already gives us a starting point for data selection, since these technologies are usually presented with a genome-wide off-target screen in their introductory publication.

Other nucleases Many promising approaches have been taken to enhance activity or specificity of CRISPR/Cas systems, an overview of which can be found in [33]. Optimising sgRNA design is only one of these avenues. Early approaches have included considering

Cas9 nucleases from other organisms such as *Staphylococcus aureus* SaCas9 [34]. Groups have used mutated versions of Cas9 in which the HNH or RuvC domain has been inactivated such that the resulting paired nickases (Cas9n) can be targeted to neighbouring sites to introduce two independent offset nicks instead of a full DSB. Another approach is to not generate DSBs but instead produce epignetic and genetic modifications by fusing with different functional domains of Cas9, a technique known as dead Cas9 (dCas9). It is used in CRISPRi and CRISPRa protocols. Not relying on the nuclease activity of the CRISPR/Cas system results in a cleaner, more predictable process since it cuts out the uncertainties connected with cleavage.

Mitigating off-target effects can further be achieved through the engineering of novel Cas enzymes such as Cas12a [35], and Cas13 [36] in the case of RNA editing. Another promising route is the engineering of high-fidelity variants of SpCas9 in which key residues known to form hydrogen bonds within the SpCas9–sgRNA–target DNA complex have been substituted, e.g. SpCas9-HF1 [37]. Enhanced Specificity SpCas9 (eSpCas9) has been developed along a similar route [38].

Whilst the vast majority of available cleavage data is still on wild-type SpCas9, making direct modelling of high-fidelity variant activity very challenging, publications such as [37, 38] offer side-by-side comparison of wild-type and high-fidelity cleavage activity under identical experimental conditions.

For the purpose of this work, we focus on the wild type Cas9 enzyme and investigate the prediction of off-targets for given guide RNA sequences.

1.1.3 Cleavage repair

Genome editing consists of DNA cleavage followed by DNA repair. There are three main intrinsic cellular repair mechanisms for double-strand breaks (see Figure 1.4):

- *Homology-directed repair* (HDR): The 5′ DNA strand is resected at the DSB, creating

a 3′ overhang. A homologous strand of a specifically introduced ssDNA repair template then anneals with the overhanging strand, resulting in a so-called *displacement loop* (D loop). Intermediate products are then resolved via DNA synthesis and ligation, leaving the DSB repaired with a part of the repair template incorporated into both DNA strands. In combination with a synthetic template, HDR can therefore be used for genome engineering. In the absence of a synthetic template, it results in error-free repair.

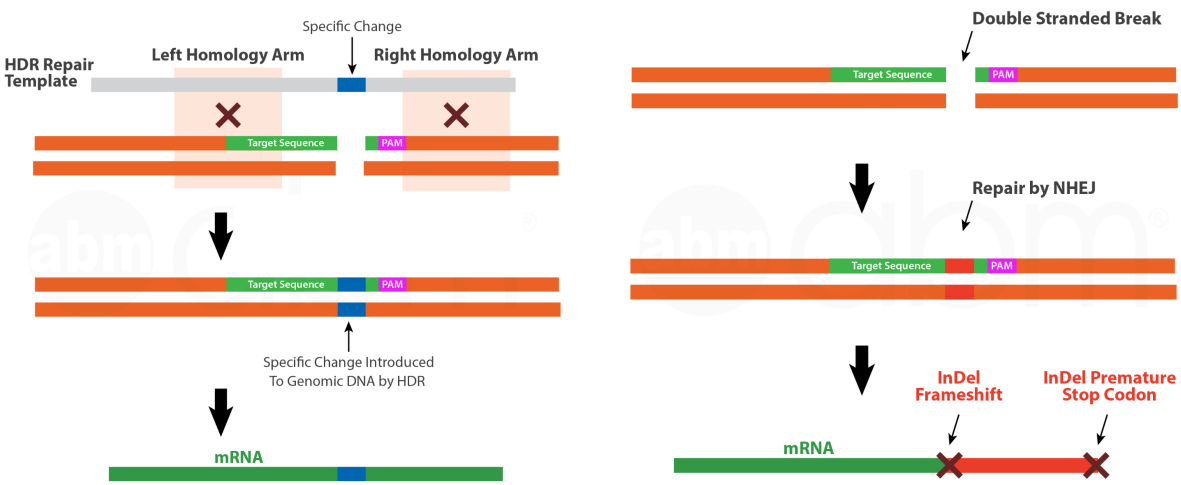
- *Microhomology-mediated end joining* (MMEJ): DNA resection around the cut site reveals homologous regions of a few base pairs on opposite strands either side of the DSB. These anneal with each other, whereupon unbound flaps are removed, gaps filled in and ligated. As a significant piece of the sequence can be removed during this process, MMEJ is an error-prone DSB repair pathway. Studies indicate that MMEJ happens with non-negligible frequency even when HDR is available [39].

- *Non-homologous end joining* (NHEJ): In the absence of a homologous repair template, a cell-internal heterodimer holds the pair of blunt dsDNA ends together, forming a *paired-end complex*. Ligases then reattach the two DNA ends together. Since no homology is required, this repair pathway is error-prone and can leave indels. These can cause a frameshift mutation in the resulting transcribed mRNA and thus possibly knock out a gene altogether.

1.1.4 DNA structure & energetics

Both the process of sgRNA–DNA heteroduplex formation and the stability of DNA itself depend on the thermodynamic properties of DNA, which we want to briefly introduce here.

As shown in Figure 1.5, DNA is a polymer molecule composed of two polynucleotide chains in a double helix arrangement. The most common form of double helical structure found in nature is a right-handed DNA helix that is called B-DNA, here simply referred to



(**a**) Homology-directed repair (HDR).

(**b**) Non-homologous end joining (NHEJ).

Figure 1.4: Means of template-free DSB repair. Images reproduced from [40].

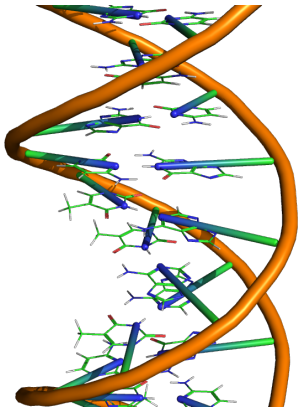


Figure 1.5: DNA double helix structure, with phosphate backbone in orange and base pairs coloured by atom type. Nucleobases are shown with individual atoms in a stick representation.

as DNA. The DNA double helix is formed of alternating deoxyribose and phosphate groups. They each carry one of four nucleobases,

- cytosine (C), thymine (T): pyrimidine bases, and
- guanine (G), adenine (A): purine bases,

which are oriented orthogonally to the helical axis, pointing inwards. Nucleotides on opposite DNA strands facing each other are called *base pairs* and appear with a periodicity of 10.5 base pairs per double helix turn, i.e. at a distance of about 3.4 Å to each other [41]. C–G and A–T are called *canonical* or *Watson–Crick base pairings*, since they are the most stable combi-

nations for all four nucleobases [42]. Between C and G, three hydrogen bonds are formed, whereas only two are formed between A and T, meaning the former require more energy to break apart than the latter. The binding free energy E_{pot} of dsDNA arises due to two main interactions:

- base pairing: hydrogen bonds within a base pair,
- base stacking: van der Waals interactions between adjacent base pairs, as well as solvent interactions [43].

A combination of these results in an energy penalty due to DNA torsion. These interactions can be used to approximate a total energy arising from atomic forces [44],

$$\begin{aligned} E_{\text{pot}}(\mathbf{r_1}, ..., \mathbf{r_N}) = &\sum_{\text{bonds}} K_r(r - r_0)^2 + \sum_{\text{bends}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\phi(1 + \cos(n\phi + \delta)) \\ &+ \sum_{i,j \in \text{nb}} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{\mathbf{r}_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{\mathbf{r}_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \right) \end{aligned} \tag{1.1}$$

where r is the length of an atomic bond, θ its bond angle and ϕ its dihedral angle which alternates between staggered and eclipsed configurations. Nonbonded interactions (nb) include a short-range Lennard-Jones potential caused by a combination of attractive van der Waals interactions and repulsion at short distances, as well as a long-range Coulomb potential due to charged electrostatic interactions between partial charges q. The former is characterised by a well depth ϵ and a van der Waals radius σ. The equilibrium parameters $r_0, \theta_0, \delta, \sigma$ are given by an atomic *force field*, e.g. AMBER [45].

1.2 Computational background

In order to both generate and establish support for hypotheses about the biological data at hand, we use techniques from the area of supervised machine learning [46]. Starting from a set $\mathcal{D}$ of N labelled observations,

$$\mathcal{D} = \{(\mathbf{x_i}, y_i)\}_{i=1}^{N}, \tag{1.2}$$

we use algorithms that provide a mapping from inputs $\mathbf{x}$ to output predictions $\hat{y}$. Traditionally, $\mathbf{x}$ is a column vector of length d which we call the number of *features*. y can be a single value, as is the case in chapter 3, or a column vector of length k which we call the number of *targets* as in chapter 4. These algorithms rely on a set of parameters that is learned through minimising a distance $l(y, \hat{y})$ between observed and predicted label. In this work, we use three classes of algorithms which are introduced below.

In contrast to this, section 3.3.3 uses an unsupervised learning algorithm in order to find a clustering of unlabelled data points $\mathbf{x_i}$.

1.2.1 Gradient boosted decision trees

In this case, the mapping algorithm has the form of an ensemble of decision trees acting on the input features, which together determine the output as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x_i}) \tag{1.3}$$

where f_k is the function representation of the k-th tree, i.e. its structure and the scores it attributes to each leaf. We call this a *random forest*. Each tree is fed a subset of the input features (*feature bagging*) for a subset of the input data points (*bootstrapping*).

Since optimising all trees simultaneously is intractable, an iterative strategy is used [47], in which we add one new tree at a time. This process of sequentially adding *weak learners* in order to achieve an overall *strong learner* is referred to as *boosting* and can be described as

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\mathbf{x_i}). \tag{1.4}$$

In this sense, instead of fitting directly to input data, the learner at iteration step t effectively fits the residuals of the sum of the predictions of learners from previous iterations. A given learner therefore depends on the performance of its predecessor, meaning learners are not independent.

In order to optimise the learner f_t at iteration step t, we minimise the objective

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t)}\right) = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x_i})\right) \tag{1.5}$$

$$\approx \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)}\right) + \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) f_t(\mathbf{x_i}) \quad + \frac{1}{2}\partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) f_t^2(\mathbf{x_i}) \tag{1.6}$$

$$= \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(\mathbf{x_i}) \quad + \frac{1}{2} h_i f_t^2(\mathbf{x_i}) \tag{1.7}$$

where we have used a Taylor expansion up to second order. We have ommitted possible regularisation terms that penalise the complexity of the resulting model to avoid overfitting. Common choices for the loss function l which we have used in subsequent chapters are

$$\text{mean squared error: } l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2, \tag{1.8}$$

$$\text{logistic regression loss: } l(y_i, \hat{y}_i) = \left[y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{\hat{y}_i}\right)\right]. \tag{1.9}$$

Equation 1.7 can be used to iteratively add branches to the tree (*grow* the tree) by evaluating the loss reduction of a given split and choosing the best split using a greedy algorithm.

Parameters for this process include the maximum depth of the tree, the minimal number of child nodes and the number of trees K in the ensemble. In chapters 3 and 4, we use an implementation of this process in form of the XGBoost package [47].

1.2.2 Neural Networks

Fully Connected Neural Network

A neural network consists of single units called *neurons* which calculate their output as a combination of their inputs. For a fully connected neural network as shown in Figure 1.6, each hidden neuron takes the inputs from all neurons in the previous layer, calculates their linear combination via matrix multiplication and feeds the result through a nonlinear *activation function*. The same goes for the output neuron which is connected to all neurons of the last hidden layer. The set of coefficients of the linear combinations for all neurons makes up the *weights* of the neural network, which characterise its behaviour.

During the training phase, these weights are optimised in order to minimise the difference between the predicted output $\hat{y}$ and the true output y as measured by a predefined *error function* in the form of equations 1.8 and 1.9. This is done by calculating the derivative of each weight with respect to the inputs. Recursively applying the chain rule for this operation results in an iterative process of optimisation (*back propagation*). New weights are calculated using a choice of *optimiser* depending on the specific network architecture, e.g. (stochastic) gradient descent, momentum, RMSprop, or Adam which is often used as a versatile method. In cases where the second order derivatives of the weights are computationally expensive to generate, we can use the conjugate gradient method (see Chapter 4.2).

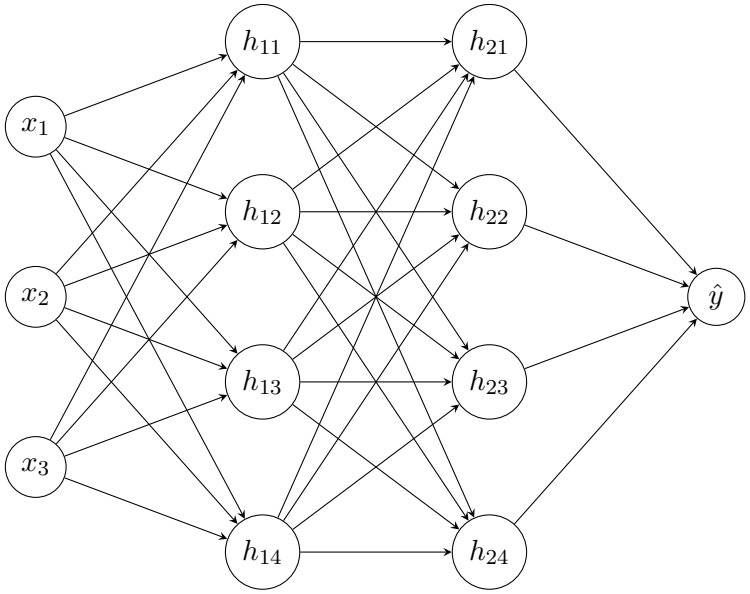


Figure 1.6: Schematic of a fully connected neural network with input vector $\mathbf{x} = (x_1, x_2, x_3)$, two hidden layers with 4 neurons each, and the scalar output $\hat{y}$.

1.2.3 Convolutional Neural Network

When working with two-dimensional data points, e.g. a one-hot encoded version of a DNA or RNA sequence, we use a similar architecture but replace the linear combination of input values with *convolutions*. In continuous space and for two functions f and g, this is defined

as

$$(f * g)(t) \coloneqq \int f(\tau)g(t-\tau)\mathrm{d}\tau. \tag{1.10}$$

For 2D convolutional networks, each neuron has a *kernel* matrix $K \in \mathbb{R}^{m\times n}$ with which the input data $I \in \mathbb{R}^{i\times j}$ is convoluted. Since this is discrete, the convolution operation boils down to

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{1.11}$$

which can be viewed as a matrix multiplication. The entries of K are learned for each neuron using the backpropagation process described above. The convolution operation results in translational invariance of the resulting network, meaning that we can discover patterns at any location within the input data.

In order to alter complexity at different layers within the network, we can use *max pool* layers which retain the maximum value within small, adjacent contiguous regions of the neuron output matrix and assemble these to a new matrix of smaller dimensionality.

Overfitting can be counteracted using *drop out* layers, which cut a random subset of neuron interconnections during a training epoch. For further reading, we refer to [48].

1.2.4 Recurrent Neural Network

In a recurrent neural network, neuron outputs are fed back into previous neurons in order for the network to achieve a memory of previous inputs. This is different from the *multilayer perceptron* architectures described above, since in recurrent architectures, information flows backwards as well as forwards within the network.

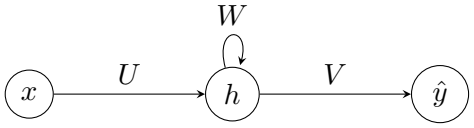


Figure 1.7: Recurrent neural network architecture.

Figure 1.7 shows the example of a simple recurrent neural network with a single hidden layer that has a recurrent connection to itself. As different iterations of the input vector $\mathbf{x}$ are

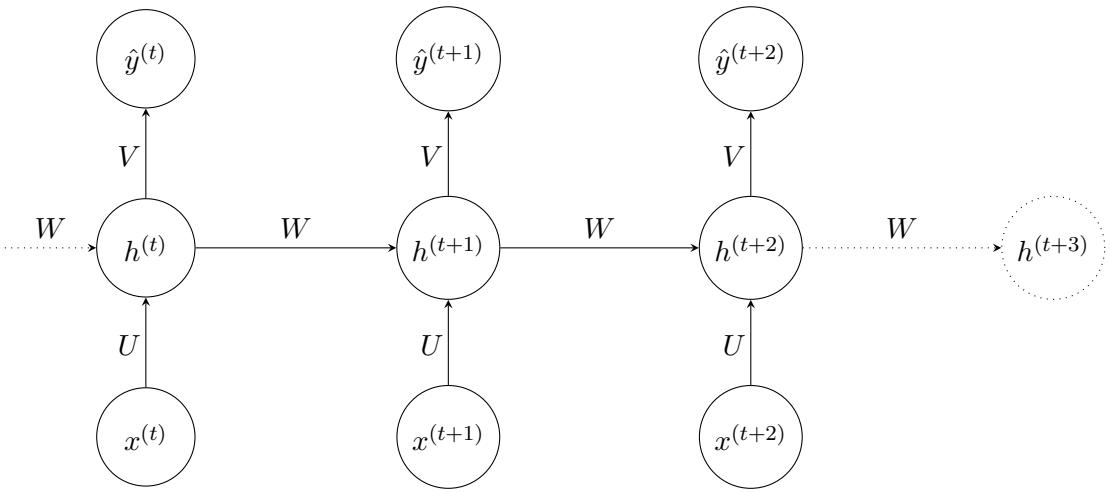


Figure 1.8: Recurrent neural network architecture from Figure 1.7. The hidden nodes $h^{(t)}$ represent the hidden states of the RNN cells at different time steps t. RNN cells take the input vector and the previous hidden state vector as input, and produce the next hidden state vector and output vector as output using the weight matrices U, V, W.

fed into the network at time step t, the hidden node gets input not only through the matrix operation U but also through the matrix W applied to the hidden output of the last time step $t-1$, which paves the way for the model to retain information from previous time steps. This is visualised in Figure 1.8, in which the simple architecture from Figure 1.7 is unfolded over time t.

Due to the long chains of multiplication over many time steps, gradient updating can easily result in vanishing gradients. This means that long-term dependencies are difficult to retain in this basic architecture [48].

A powerful class of RNNs that adresses this problem are *gated RNNs* which consist of self-loops that are controlled (gated) by another hidden unit. Two popular architectures for this are *long short-term memory* networks and *gated recurrent units* (GRUs). These have the aim of keeping the gradient flow for long durations without "forgetting", i.e. losing meaningful long-term dependencies. An example of how a GRU architecture determines the hidden state $h^{(t)}$ from the previous hidden state $h^{(t-1)}$ and the input vector $x^{(t)}$ is shown in Figure 1.9. An update gate represented by z determines ("gates") what degree of past information encoded

in the hidden state $h^{(t-1)}$ is to be carried forward to the next time step t. This is done using a separate set of weight matrices W_z, U_z. The reset gate r is responsible for determining the importance of the previous hidden state $h^{(t-1)}$ and which proportion of it is used to calculate the new hidden state $h^{(t)}$. The latter is finally calculated as a weighted average of the previous hidden state $h^{(t-1)}$ and the activation vector $\tilde{h}^{(t)}$ containing the effect of the reset variable $r^{(t)}$, where the weight is given by the update variable $z^{(t)}$. To summarise in equation form,

$$\text{update gate:} \qquad z^{(t)} = \sigma\left(W_z x^{(t)} + U_z h^{(t-1)}\right), \tag{1.12}$$

$$\text{reset gate:} \qquad r^{(t)} = \sigma\left(W_r x^{(t)} + U_r h^{(t-1)}\right), \tag{1.13}$$

$$\text{activation vector:} \qquad \tilde{h}^{(t)} = \tanh\left(W_h x^{(t)} + U_h\left(r^{(t)} * h^{(t-1)}\right)\right), \tag{1.14}$$

$$\text{next hidden state:} \qquad h^{(t)} = \left(1 - z^{(t)}\right) * h^{(t-1)} + z^{(t)} * \tilde{h}^{(t)} \tag{1.15}$$

where $\sigma(x) \coloneqq 1/(1+\exp(-x))$ is the *logistic* or *sigmoid function*, and $*$ denotes the *Hadamard product* (element-wise matrix product).

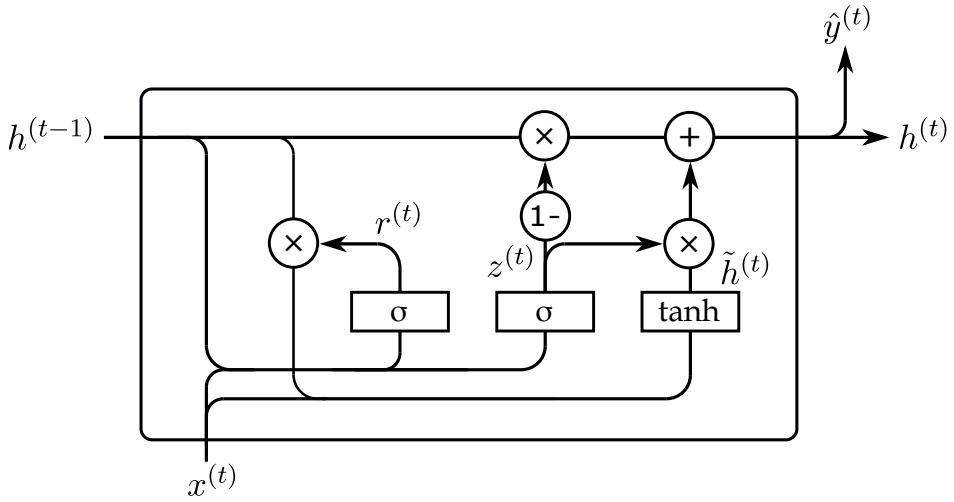


Figure 1.9: Gated Recurrent Unit as characterised by equations 1.12 - 1.15, showing the update gate of vector $z^{(t)}$ and reset gate of vector $r^{(t)}$.

GRUs can be trained bidirectionally, i.e. with t running down- as well as upwards, such that sequential context can be learned in both directions. To do this, two GRU layers h_{forward} and h_{backward} are stacked, with their interconnections running in opposite directions. For a visualisation of this, see [48]. This means that more intricate dependencies along the time

(e.g. DNA sequence) axis can be learned than in the local, low-parameter case of convolutional kernels discussed above.

1.2.5 MOSAICS

For our structural calculations, we use the computational framework MOSAICS (Methodologies for Optimization and Sampling in Computational Studies, [49]). It is a powerful *in silico* tool that is capable of sampling high-dimensional conformational spaces for RNA/DNA molecules based on energy functions defined by atomistic force fields [45]. Hierarchical natural moves can be utilised to efficiently sample along axes that coincide with molecular degrees of freedom. Sampling can be driven via various procedures including Markov chain Monte Carlo and conjugate gradient sampling, depending on the complexity of the molecule and the desired accuracy.

1.3 Existing CRISPR modelling efforts

1.3.1 Cleavage data repositories

As suggested in [50], the chief bottleneck in "understanding and modelling biological systems is shifting from data collection to data analysis and integration". In this sense, the vibrant CRISPR community has developed a variety of online tools to aid with guide design, finding suitable CRISPR arrays, etc. A non-comprehensive overview of this can be found in Table 1.2. Most of these databases act as higher-level user interfaces for raw and processed experimental data from repositories such as GEO [55], ENCODE [56], BioProject/BioSample [57], SCREEN [58] and BioGRID [59].

1.3.2 sgRNA specificity modelling efforts

Due to the fast-moving nature of this computational field, it is challenging to obtain a complete overview, let alone direct comparison of existing computational tools. Between refer-

task	examples
find CRISPR loci on prokaryote genomes	[51, 52]
computationally finding possible off-targets	[53]
cleavage data with only on-target resolution	[54]
cleavage data with off-target resolution	n/a

Table 1.2: Non-comprehensive overview of online databases related to CRISPR. They usually fall into the first three categories, with a clear lack for a service offering annotated cleavage data with off-target resolution. This is where we want to contribute an online platform.

ences [33, 60, 61, 62], we can approach a comprehensive overview of current computational publications.

A timeline of sgRNA design tools can be found in Figure 1.10, whereas Table 1.3 gives an overview of computational tools for prediction of sgRNA specificity.

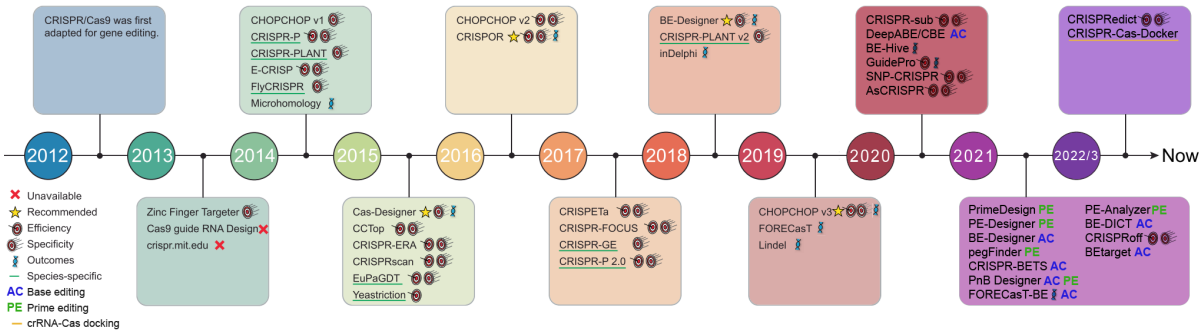


Figure 1.10: Timeline of web-based tools for sgRNA design (selection). Until 2019, this is a reproduction of Figure 2 from [61].

Off-target prediction requires the identification of putative off-targets along the genome, which is usually done using an existing search tool such as BWA [71], Bowtie [72] or Cas-OFFinder [73]. These tools find all genomic loci with $\leq k$ mismatches where k is usually between 5 and 8. However, experimental data shows that off-target effects with $k \geq 10$ can occur with non-negligible frequency. Accurate and comprehensive identification of putative off-targets within a given reference genome requires different methods such as bi-directional alignment [74].

Due to the relative scarcity of off-target datasets at base pair resolution compared to on-target datasets, innovations in the field of off-target prediction have mainly been due to the

Tool	Enzymes	Methods	Main features	Quantitative metrics
CasOT [108]	Cas9	alignment	unlimited mismatch number, paired-gRNA mode, annotation	slow
Cas-OFFinder [8]	custom	alignment	unlimited mismatch number, GPU acceleration, web support	middle, fast (use GPU)
sgRNAcas9 [110]	Cas9	alignment	max 5 mismatches, paired-gRNA mode, annotation, risk evaluation	slow
FlashFry [74]	custom	alignment	unlimited mismatch number, multiple on/off-target scores, annotation	fast
Crisflash [43]	Cas9	alignment	unlimited mismatch number, variant data support	fast
MIT [41]	Cas9	scoring	20 bp sgRNA without PAM	ROC-AUC = 0.87, from [36]
CCTop [93]	Cas9, Cpf1	scoring	empirically score based on number of mismatches	ROC-AUC = 0.77, from [36]
CFD [23]	Cas9	scoring	20 bp sgRNA with PAM (enable non-canonical PAM)	ROC-AUC = 0.91, from [36]
CRISPRoff [4]	Cas9	scoring	energetics property and sequences composition	ROC-AUC = 0.98
uCRISPR [119]	Cas9	scoring	energetics property and sequences composition	Pearson correlation = 0.75
CRISTA [1]	Cas9	scoring	machine learning, sequences composition and epigenetic feature	ROC-AUC = 0.96
Elevation [62]	Cas9	scoring	machine learning, integrate both CFD model and epigenetic features	ROC-AUC = 0.98
DeepCRISPR [18]	Cas9	scoring	deep learning, sequences composition and epigenetic feature	ROC-AUC = 0.98
AttnToMismatch [63]	Cas9, Cas12a	scoring	attention-based transformer applied to sequence embeddings	
dl-offtarget [64]	Cas9	scoring	lossless, explicit encoding of sgRNA and DNA, no epigenetics	
CnnCrispr [65]	Cas9	scoring	sequence embedding using GloVe model, biLSTM	
Lin *et al.* [66]	Cas9	scoring	OR encoding of sgRNA and DNA means information loss	
CRISPR-Net [67]	Cas9	scoring	lossless sequence encoding using two additional features, takes indels into account	
Lazzarrotto *et al.* [31]	Cas9	scoring	gradient boosting applied to single large, uniform dataset	
CRISPR-IP [68]	Cas9	scoring	lossless sequence encoding plus additional function channel	
MOFF [69]	Cas9	scoring	split mismatch tolerance into guide-intrinsic and target-specific contributions	
CrisprDNT [70]	Cas9	scoring	transformer model using generalised cross entropy loss for label noise suppression	

Table 1.3: Computational methods for prediction of sgRNA specificity. Publications until 2020 are given as a reproduction of Table 2 from [61].

application of novel modelling techniques to combinations of existing datasets. Milestones were Elevation [60], utilising a linear regression model which takes both sequence and chromatin accessibility features into account, as well as DeepCRISPR [75] which successfully applied a convolutional neural network (CNN) to sequence and epigenetic data without performing explicit feature engineering. A systematic comparison of several fully connected and convolutional architectures was done in [66] and showed that among these, a convolutional neural network with batch normalisation and max-pooling layers yielded the highest area under receiver operating curve (AU-ROC). Recently, attention-based networks [63] and recurrent neural networks [68] have been applied successfully to this problem.

An example of a novel dataset driving prediction performance is reference [31] where the authors applied an established gradient tree boosting algorithm to their novel dataset, which was an order of magnitude larger in terms of measured off-targets than previous datasets. It should be noted that demonstrating model generalisation and comparison with existing predictors is a significant challenge in this approach.

1.3.3 Existing editing outcome prediction algorithms

It used to be a common notion that mutations resulting from Cas9-induced DSBs are randomly generated by the aforementioned NHEJ and MMEJ pathways. Several publications have since demonstrated that these mutations are indeed nonrandom and could be predicted using sequence properties [76, 77, 78]. Most of the models mentioned in Table 1.4 rely on purely sequence-based features such as base pair identities, dinucleotides or microhomology features.

name of corresponding web service	year	main modelling technique	ref.
FORECasT	2018	multinomial logistic regression	[76]
inDelphi	2018	NN, multitask framework learning	[77]
Lindel	2019	logistic regression (LR)	[78]
SPROUT	2019	gradient boosted tree ensembles	[79]
Apindel	2022	BiLSTM RNN	[80]
PREDICT (*prime editing*)	2022	bidirectional RNN	[81]

Table 1.4: Overview of CRISPR editing outcome predictors published to date.

FORECasT chooses a shallow learning approach, using a logistic regression function to model the probability of each possible outcome, which is fitted to experimental data via minimisation of the L_2-regularised KL divergence between the two probability distributions. inDelphi uses two separate neural networks which predict probabilities of microhomology-driven and non-homologous indels, respectively. BE-Hive [82] should be mentioned here as a predictor for base editing outcomes. It uses a twofold structure wherein CRISPR-caused bystander edits (edited C or A nucleotides within the targeting window) are predicted differently from the edit efficiency at the desired target side. None of the models above take structural free energy calculations into account.

In order to increase visibility and accessibility, most of the above models come with web interfaces which allow the extraction of model results for single sequences and DSB points. An example of this is shown in Figure 1.11. Using these, we noticed that the relative frequency of indel outcomes appears to be overemphasised for indels whose length is an integer

multiple of 10, which is close to the number of base pairs per turn in the DNA double helix. This sparked our interest in indel explanation using structural calculations (section 4.2).

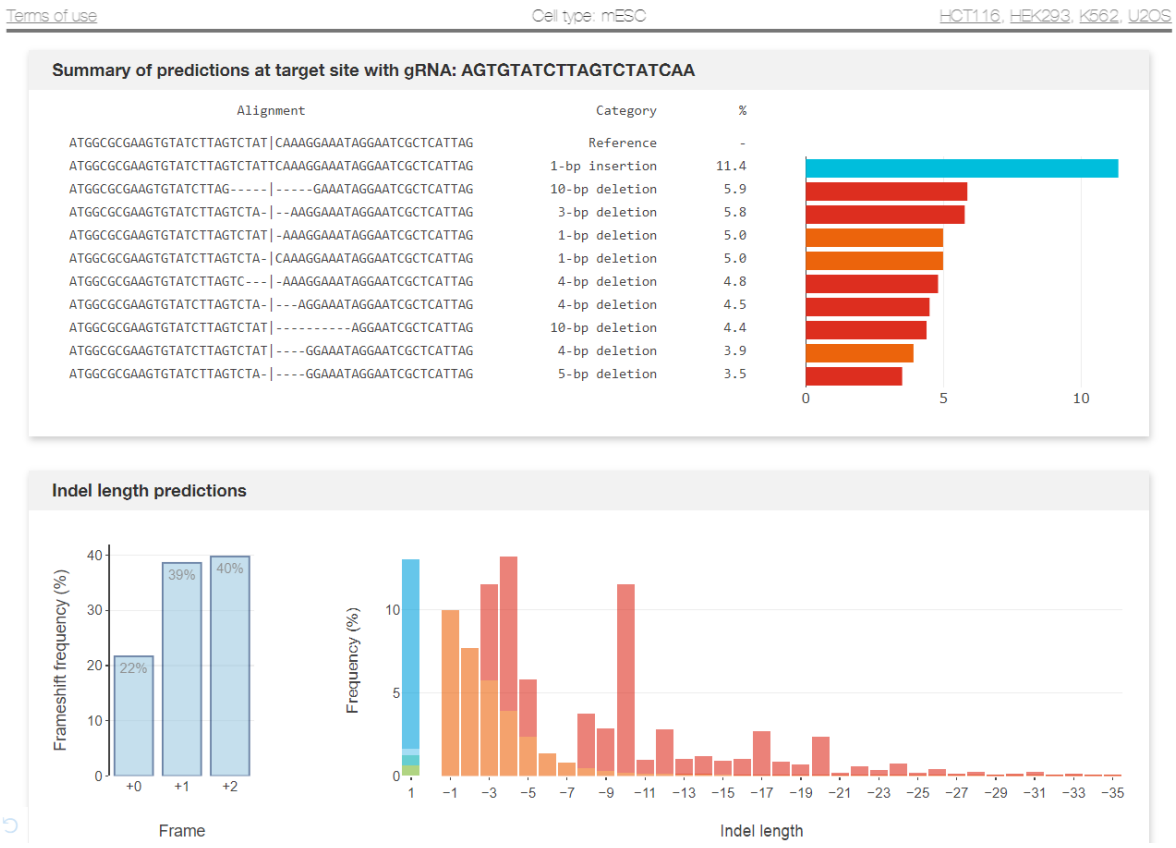


Figure 1.11: Output of the inDelphi web service [77] for a certain example sequence. The seemingly overemphasised peaks around indel lengths of 10 and 20 sparked our investigation into structural calculations for interpretation of editing outcomes.

1.3.4 Industry review

Next to a variety of companies providing reagents, companies such as Synthego offer services described as "full stack genome engineering". This includes tasks all along the genome engineering pipeline, such as a sgRNA optimiser for gene knockout, bespoke reagents including synthetic RNA as well as sequencing analysis tools. It surpasses the offers from competitors such as GE Healthcare in terms of completeness. A key contribution is the use of cloud-based computational resources instead of relying on local computing power. Whilst the sgRNA design tool appears to be rather sophisticated, its corporate ownership means that no design

choices are published.

1.3.5 Conclusion

Based on the range of different training and validation data sets which are used by existing models, we have decided to dedicate the next part of this work to the creation of a comprehensive, computationally efficient and well-referenced bespoke dataset containing all available data of necessary quality and resolution. This can then be used to train cleavage prediction models, as well as inform gene editing research as a whole.

Chapter 2

Database platform for CRISPR/Cas9 off-target cleavage assays

This section has been published in *Nucleic Acids Research* as "crisprSQL: a novel database platform for CRISPR/Cas off-target cleavage assays" [83]. The supplementary information given there has been embedded in the text.

In addition to providing a training and validation dataset for our cleavage modelling efforts, we want to create a publicly accessible, interactive online version of our dataset for the benefit of the wider research community. We envision that it will be used by computational and biological researchers alike, aiding in tasks such as guide design and off-target cleavage modelling.

2.1 Introduction

Several clinical applications of the CRISPR/Cas gene editing technology have been realised to date, such as personalised cancer treatment with genetically modified chimeric antigen receptor (CAR) T cell-based therapies [9]. A variety of possible therapeutic applications in the fields of genetic diseases, infectious diseases and cancers are being investigated [10], as exem-

plified by 38 clinical studies referring to CRISPR currently registered on http://clinicaltrials.gov (accessed 07 August 2020). As Dai *et al.* state, "one of the major hurdles to the clinical translation of CRISPR/Cas9 is its off-target effects" [8], referring to the unwanted nucleotide insertions or deletions created by CRISPR outside of the targeted genomic locus. Electing an appropriate guide RNA (gRNA) for the desired DNA target sequence is considered a "crucial first step in avoiding off-target effects" [84]. In order to mitigate off-target effects, "a set of validated off-target sites should be compiled and used to create a reagent set to quantitatively study off-target effects via NGS methods" [85]. With off-target data spread across supplementary data of various publications, this remains a tedious and time-consuming task.

Besides the design and execution of medical studies, a comprehensive database of well-known off-target interactions is required for the validation of novel off-target detection methods [27] and new implementations of existing methods in a laboratory setting.

The development of computational off-target prediction algorithms [65, 75] also relies on the presence of a comprehensive dataset quantifying off-target effects of gRNAs on DNA from certain cell lines, which is usually annotated with a selection of epigenetic markers by the respective authors. The varying choice of annotation sources, training data and test procedures can result in issues of reproducibility [86].

We contribute an online database which promises to enhance clinical, laboratory and computational research by providing the first one-stop source of off-target resolved CRISPR/Cas cleavage data to date, accompanied by epigenetic annotations and visualisations. In order to provide maximum value for a range of fields including transcriptomics, gene knockout experiments, editing safety-driven guide design and cleavage efficiency prediction, crisprSQL offers sequence-resolved data whilst also bridging the gap to gene resolution by attaching GENCODE gene names to interaction targets.

2.2 Materials and Methods

2.2.1 Data Acquisition

Cleavage rate sources & formats There have been numerous publications scrutinising the off-target effects of *S. pyogenes* CRISPR/Cas9 over the past decade (see Table 2.1), most of which rely on second-generation whole-genome sequencing tools. First-generation approaches such as the T7 endonuclease assay only find off-targets among a previously *in silico* defined set, therefore only yielding limited insight into the full off-target profile.

In order to provide the most detailed insight into cleavage processes, we have chosen to include only those studies which offer base pair-resolved on- and off-targets, as well as an unambiguous assignment of which specific guide caused a given cleavage event. This excludes pure on-target studies, pure gene knock-out studies and most studies done on whole gRNA libraries. To our best knowledge, such a database has not been realised before. We hypothesise that it is mainly the guide-target interface and its biological environment that drive CRISPR cleavage. We therefore intend on building models that have access to information about both sgRNA and (off-)target site, even if this limits us to a certain range of study formats.

The use of ENCODE Tier 1 immortalised cell lines [87] facilitates comparison between experiments and annotation with cell-specific metadata. We rely on the assumption that sequence concordance between cell lines of the same name at different laboratories is near-perfect, as was verified in [88]. There, the authors assert that "sequence variances are minimal" between HeLa cells of 14 strands taken from 13 international laboratories. Sequence concordance between pairs of strands, defined as the proportion of the overlapping variant calls compared to a HeLa reference that are consistent between two alleles on either strand, ranged from 89.6% to 95.1%. The number of gene copies (gene dosage), however, can differ significantly between strands and cause differences in gene expression [88, 89]. This is beyond the sequence-focused scope of our database.

ref.	technique	detects	guides	targets	annot.	assembly	cell lines
[90]	T7E1	heterodupl. DNA	9	88	0	hg19	U2OS, HEK293, K562
[91]	targeted PCR	genome change	10	116	11	hg19	K562
[27]	Guide-seq	DSBs	12	575	381	hg19	U2OS, HEK293
[21]	Digenome-seq	genome change	2	162	7	hg19	HAP1, K562
[92]	BLESS, ChIP	DSBs	6	87	53	hg19, mm9	293FT, N2a
[93]	HTGTS	DNA junctions	3	87	87	hg19	HEK293T
[94]	IDLVs	viral integration	1	13	13	hg19	HEK293T
[95]	Digenome-seq	genome change	10	258	234	hg19	HeLa
[37]	Guide-seq	DSBs	7	61	28	hg19	U2OS
[38]	BLESS	DSBs	3	31	31		HEK
[30]	CIRCLE-seq	DSBs	18	7374	3493	hg19	HEK293, U2OS, K562
[96]	Guide-seq	DSBs	5	203	107	hg19	U2OS
[29]	SITE-Seq	DSBs	8	1630	210	hg38	HEK293
[97]	DIG-seq	genome change	7	141	132	hg19	HeLa
[98]	Guide-seq, WGS	DSBs	31	426	0	mm9, rn5	mouse & rat embryos
[60]	Guide-seq	DSBs	10	272	160	hg19	U2OS
[99]	NucleaSeq	cleaved products	2	14108	0		custom DNA library
		sum	144	25632	4947		

Table 2.1: Data sources included in crisprSQL (version 10 July 2020), chronologically ordered by publication date. With 25,632 guide-target pairs, our dataset is more than an order of magnitude larger than both the CRISPOR [100] and DeepCRISPR [75] off-target datasets with less than 680 guide-target pairs each. Note that we have excluded guides with less than two reported off-targets, but kept GC-rich guides and guide-target pairs with low cleavage frequencies. The 'annotated' column counts how many of the respective guide-target pairs are annotated with at least one epigenetic marker. The term 'target' includes both on-targets, i.e. the intended cutting site which is homologous to the guide sequence, and mismatched off-targets.

We manually extracted the cleavage frequencies from the references in Table 2.1 using an online PDF-to-csv tool wherever these were not supplied in text form. Where mutation frequencies, indel frequencies or similar terms were given, we used these values for the cleavage frequency value. In case only relative read counts were given, cleavage frequency was calculated as the read ratio relative to the sum of reads for a given guide. The included publication [97] chose identical target sequences in both open and closed chromatin regions, thereby promising very telling data on the effect of chromatin accessibility. Publication [101] which is included in the CRISPOR dataset [100] was excluded here because extraction of the

relevant data (gRNA, target locus or sequence, binding & cleavage frequency) was not unambiguously possible. For some studies, we additionally obtained unpublished data points from the same measurement series which authors were happy to share on request.

The off-target cleavage assays included in crisprSQL can be divided into two categories. DSB capture approaches in a cellular context (*in situ*) are more physiologically relevant, albeit prone to false positives due to cellular processes which may generate non-Cas9 related DSBs. IDLV [94] and GUIDE-seq [27] rely on the non-homologous end joining-mediated integration of sequence markers into sites upon DSB. Typical sensitivities range from 0.1% [27, 93] to 1% [94]. Other *in situ* techniques include HTGTS [93] (translocation of DSBs to bait DSBs) and BLESS [92] (integration of biotinylated linkers). IDLV integration does not always happen at the exact DSB location, whilst GUIDE-seq is limited to blunt-end DSBs [102]. Whilst a blunt-ended DSB is the predominant outcome of Cas9, it has been shown that Cas9 can generate 1bp overhangs at the DSB [103] which are then not detected by GUIDE-seq. The same is true for Digenome-seq [21] due to computational constraints in its initial version [104].

In vitro approaches outside of a cellular context do not include effects of epigenetic factors and chromatin on DSB formation, making the subsequent validation of observed off-targets in cells desirable. SITE-Seq [29] has been shown to overestimate off-targets relative to those observed in a cellular context. CIRCLE-seq [30] shows high sensitivity through enrichment of Cas9-cleaved genomic DNA, which leads to a considerable level of background noise overshadowing true positive off-targets [102]. DIG-seq [97] uses chromatin-associated DNA and can therefore assess the influence of chromatin on Cas9 cleavage. Experimental protocols for the different assays entail different relative concentrations of Cas9 expression plasmid (100 ng–1 µg), gRNA-encoding plasmid (50 ng–1 µg), and possible additions such as blasticidin resistance-inducing plasmids. The ratio of Cas9 plasmid : gRNA plasmid ranges from 1:1 to 3:1, and the time between Cas9 transfection and cell harvest for DSB detection ranges from

24 h [92] to 5 days [98].

With the exception of the NucleaSeq assay [99], crisprSQL only includes data points which have been validated in cells. For *in vitro* methods, this is a certain fraction of DSB sites identified by the respective assay, ranging from 62% [98] to 100% [21, 29]. For these methods, validation data in cellular context has been gained through targeted deep sequencing.

It has become convention to give guide and target sequences as their respective complement in the protospacer strand such that a canonically matched guide-target pair shows as identical base letters. To extract the actual binding sequences from our database, one therefore has to exchange thymine for uracil in the saved gRNA sequence, and take the canonical complement of the saved target sequence.

Data processing The resulting csv file was imported in Python using `pandas`. Missing genomic locus information was filled in using the `blastn` tool. For sequences which were shorter than 23 bp, `samtools faidx` was used to extend them to this length, making sure the found sequence contains at least the 10 PAM-proximal base pairs of the original, has no uncalled bases (N) and an identical PAM. It has been reported that sequence context plays a significant role for cleavage efficiency [105, 100]. We therefore provide the 169 bp sequence context around the centre of each target obtained using `samtools faidx`. This length has been chosen such that the nucleosome core DNA length of 147 bp can be retrieved around any of the 23 bp of each target. To annotate targets with the matching gene name, we retrieve the gff3 files for the respective genomes from the GENCODE database [106], extract the gene annotations and convert them to bed format. Overlap is then checked for each target sequence.

Delivery method In order to be able to scrutinise the effect of the delivery method of the CRISPR machinery into cells, the database contains a column characterising the mode of delivery (lipofection/electroporation/other), and a column indicating whether a given data point was gained in a cellular context via an unbiased (genome-wide) detection method.

Epigenetic factors Binding and cleavage as natural events do not only depend on sequence information as acquired in the first step, but also on further atomistic parameters of the guide–target heteroduplex. These epigenetic factors have been obtained experimentally using a variety of techniques. The ENCODE database [107] offers a platform which holds this experimental data and allows to search it. SCREEN is a search tool for the ENCODE database which summarises the presence of certain epigenetic factors / candidate regulatory elements along the human (hg19) and mouse (mm10) genome, for several cell lines or tissues. We obtained bed files from the SCREEN web interface which contain genomic regions in which epigenetic markers are present for four individual markers. This choice has become customary in the field of computational off-target prediction [63, 75, 108]: (a) CTCF: CCCTC-binding factor, present at sites which activate gene transcription and are related to chromatin organisation [109], (b) DNase-seq: sites sensitive to the DNase I enzyme correspond to open chromatin (i.e. accessibility of the DNA strand), (c) RRBS: DNA methylation state, obtained by converting unmethylated cytosines to uracil, (d) H3K4me3: trimethylation of the lysine-4 on the histone H3 protein, associated with transcription of nearby genes [110]. A full list of the assay files which we used can be found at `http://crisprsql.com/epigen.php`. We only considered the regions where the marker value is in the 95th percentile across the genome by choosing a cutoff of $Z \geq 1.64$ on the SCREEN web page. Genomic regions present in our database which overlap with a marked region in one of these bed files are annotated with the respective Z score, as well as the respective SCREEN accession string to ensure traceability. Overlap is assessed using the `bedtools intersect` command.

Since SCREEN only offers this data for the hg19 and mm10 gene assemblies and a selection of cell lines, data for other reference genomes was obtained from the ENCODE database itself. The fifth epigenetic marker came from the DRIP-seq assay [111], which detects R-loops between DNA and RNA formed during transcription, exposing single-stranded DNA and thereby possibly causing genome instability. Data files for the DRIP assay had to be

converted from bigWig to bed using the BEDOPS suite [112]. We downloaded further data gained through the DRIP assay in bed format from the publications [111, 113, 114, 115] via the GEO database [116]. Target loci for which an epigenetic overlap has been found using BamTools [117] were updated using the quality score of the respective bam file region(s) and the unique ENCODE accession string of the bam file. Where two or more overlaps were found for a given target region, only the maximum quality score was kept.

We should note that not all epigenetic markers are available for every cell type and genome assembly (see Table 2.2). As a workaround, sequences could be realigned to other genome assemblies within species (i.e. hg38 to hg19, mm9 to mm10). As this might introduce impurity into the data, it has not been implemented. However, with the data from Table 2.2, we are able to check annotation for more than 97% of the off-target loci with at least two of five epigenetic markers.

assembly	cell line	epigenetic marker				
		DNase	CTCF	H3K4me3	RRBS	DRIP
hg19	HEK293	✓	✓	✓	✓	✓
	K562	✓	✓	✓	✓	✓
	HeLa	✓	✓	✓	✓	✓
	U2OS	✓	✓	✓		✓
	HAP1	✓	✓	✓		
hg38	HEK293	✓	✓	✓		
	K562	✓	✓	✓		
	HeLa	✓	✓	✓		
	U2OS					
	HAP1	✓				
mm10	embryonic tissue	✓		✓		
mm9	N2a					
rn5	embryonic tissue					

Table 2.2: Availability of epigenetic markers by cell line in the ENCODE, SCREEN [107] and GEO [116] databases.

Interaction energies In reference [13], the authors present an approximate binding energy model (termed CRISPRspec) for the Cas9–gRNA–DNA complex, including contributions from the gRNA–DNA hybridisation and the opening of the DNA–DNA duplex in the

Database field	GU_allowed	pos_weight	pam_corr	grna_folding	dna_opening	dna_pos_wgh
energy_1	False	False	False	True	False	False
energy_2	False	True	True	False	True	True
energy_3	False	True	False	False	True	True
energy_4	False	False	True	False	False	False
energy_5	False	False	False	False	False	True

Table 2.3: Arguments for the crisprspec.get_eng() function with which CRISPRspec energy values have been retrieved.

target region. RNAfold [118] is used to calculate the approximate free energy of the gRNA folding based on sequence information. We calculate a representative choice of five energy contributions for each gRNA–target sequence pair and annotate the respective data point. The function arguments used to retrieve these can be found in Table 2.3.

2.2.2 Data storage

In order to achieve high enough read and write speeds, we use a Python-implemented SQLite database in local memory for data storage and augmentation. The database is internally separated into three data tables which contain information about the included cleavage assay publications, epigenetics studies datasets and measured guide-target cleavage data points, respectively. This ensures that we are able to save all necessary information in a structured and efficient way which allows full traceability (see Figure 2.1). From this dataset, we are able to extract convenient views, e.g. in csv format for import into third party software, or in MySQL format to provide a website interface.

In conjunction with external tools, our database can be used to suggest optimal guides for cleavage at a certain locus whilst minimising off-target effects.

2.2.3 Website interface

Implementation In order to make our database easily accessible, we implemented a web-interface using a MySQL-based instance of our database at http://www.crisprsql.com. The

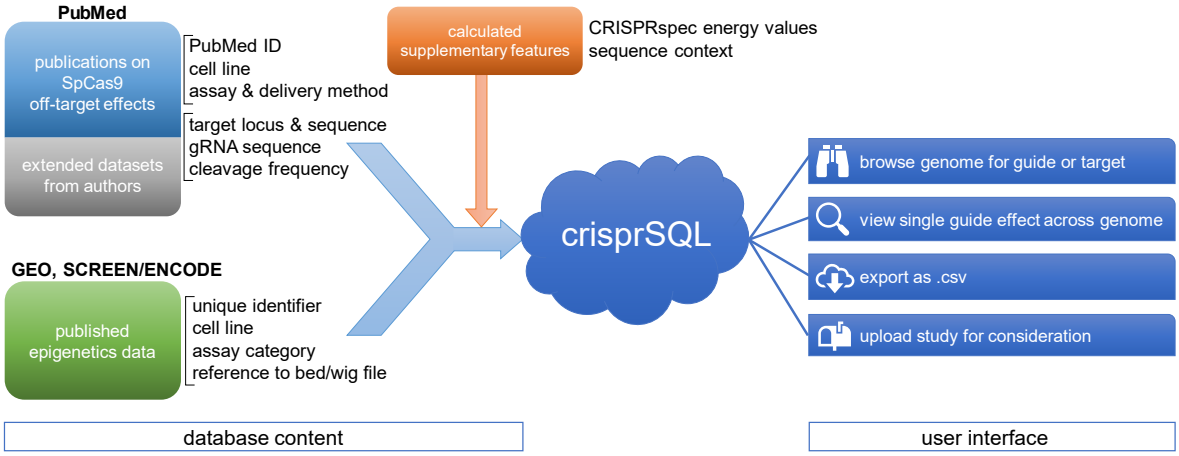


Figure 2.1: Overview of the crisprSQL database. Cleavage and epigenetics data (left blue/green fields) are joined together and supplemented by attributes calculated from their properties (top orange field). Additional cleavage data obtained from the authors was included where possible. Brackets indicate single data tables. This forms the crisprSQL database, which has an online user interface supporting data browsing, export and upload of new studies (right).

frontend and search function are rendered using PHP.

An appropriate subset of database columns was chosen for this so as to not overload the user interface. Columns describing the CRISPRspec energy contributions and the sequence context are not included in the web interface, only in the csv file. Furthermore, the website shows only the presence or absence of a nonzero epigenetic score but not its value. The web interface offers a full-text search for specific guides, target regions or cell lines, and a download of the full database in csv format. Similar services already exist for high-throughput on-target studies [54], but to our best knowledge, crisprSQL is the first online database for Cas9 off-target assays.

Through use of a state-of-the-art front-end framework, the website is fully functional on phone and tablet screens as well. An overview of the included off-target studies shows the respective detection assays, number of involved guides and targets as well as the gene names which have been predominantly cleaved (Figure 2.2). It is possible to browse through all included guide sequences grouped by sequence and cell line (Figure 2.3) as well as search for

guide or target sequences, GENCODE gene names or loci. Search results are subsequently shown in a table linking to the original publications, to a genome browser showing the vicinity of the cut site as well as to studies which have demonstrated epigenetic markers at the respective DNA loci. Hyperlinks to all involved study publications and epigenetic data repositories ensure transparency and precise traceability of the included data. The search result is visualised as a barcode-type histogram plot depicting the locations and observed cleavage rates of found targets across the genome. A similar barcode is given for the editing profile of each guide, allowing a first insight into its specificity. The database therefore lends itself as a powerful in-depth literature research tool for planning off-target studies, such as comparing known off-target performances of specific guides or finding guides whose off-target effects include a specific gene or locus.

The website also allows submission of off-target study results as text files containing guide-target pairs, cell line, measured cleavage rates and publication reference which will then be processed and integrated into the database. In this way we hope to contribute to data visibility and accessibility in the genome engineering field, as well as model comparability and reproducibility in the area of off-target prediction algorithms.

2.3 Results and Discussion

We have established a data processing pipeline and database tool which fill the niche of a thoroughly curated and annotated collection of base-pair resolved CRISPR/Cas9 off-target cleavage assay data. Besides providing a comprehensive search tool for gRNA design on both base pair and gene level, it can act as a data source for both the experimental validation of off-target detection pipelines and for computational off-target cleavage prediction algorithms. This will support experimental research into gRNA design for a variety of applications in the wider gene editing and transcriptomics fields, as well as enhance transparency and reproducibility of computational studies.

The quality of our data processing pipeline can be assessed qualitatively along several dimensions:

- completeness: After a thorough literature search, we included all matching data sources which were available at the time. As detailed above, we imposed strict selection criteria to ensure consistent data formats.

- robustness: Our annotation pipeline is robust to missing input fields. In the processing step, missing fields, e.g. loci or sequences, were retrieved from genome assemblies.

- fidelity in reproducing the original data: We cross-checked randomly selected data points against their original sources.

- transparency: The process of data manipulation is given in detail above and in our publication, which we refer to on the website.

- traceability to original sources: At every stage of abstraction, we give the hyperlink to the publication that details how the respective point has been gained experimentally.

- accessibility: Through use of the bootstrap framework, the website is optimised for various viewers. Examples for database queries are given for all parameters, as well as histogram visualisations for the distributions of (off-)targets across the genome.

- usability across application domains: The addition of gene IDs makes the underlying off-target datasets usable for transcriptomics applications.

- scalability: We have not tested the pipeline with an order of magnitude more data points. Given that the annotation process currently takes around three hours on a desktop CPU and is highly parallelisable, we conclude that scalability is not an issue yet.

As noted above, experimental protocols for the different assays differ considerably in the concentration ratios of Cas9- and gRNA-encoding plasmids as well as the timespan between

cell transfection and harvest. GUIDE-seq [27], Digenome-seq [21] and CIRCLE-seq [30] require a considerably more intricate computational pipeline than other methods to extract enriched genomic intervals after alignment to a reference genome and check for nuclease-induced edit sites. These represent specific assumptions and normalisation steps, creating the need to further investigate when absolute cleavage frequency values are compared between studies, and to at least normalise the cleavage rate distribution for each single study with a monotone, nonlinear function when relative rankings of cleavage frequencies of a cross-study dataset are desired. References [60] and [119] suggest transforming cleavage rates to approximate a Gaussian with zero mean and variance $\sigma^2 = 2$ using the nonlinear Box–Cox transformation [120]. Values above and below $\pm 2\sigma$ should be capped in order to achieve a fixed value range. As a logarithmic transformation, it is only valid for values larger than zero, which is why measured cleavage rates below the lowest reported measurement accuracy (10^{-5} in our case) should be manually transformed to -2σ (this should only apply to a select few data points). An example of such a normalisation is shown in Figure 2.4.

References [31, 32] shows that despite differences in experimental protocol, different off-target cleavage profiling techniques can result in highly overlapping cleavage distributions (Wilcoxon $p = 0.52$ between GUIDE-seq and Extru-seq, $R^2 = 0.8296$ between CIRCLE-seq and CHANGE-seq). This supports our approach of combining studies with differing experimental protocols.

Even though Digenome-seq and Extru-seq share large parts of their experimental protocols, they yielded largely incompatible cleavage distributions in reference [32]. The authors attribute this to the large number of false positives in the Digenome-seq distribution. The at times incomplete validation of *in vitro* off-target effects in cells further complicates the comparison of cleavage sites between studies. For a direct comparison of sets of cleavage assays regarding their relative performance, we refer to references [30, 95, 97].

2.3.1 Future Developments

We invite submissions of appropriately resolved cleavage frequency data by experimental authors, which will be run through our annotation pipeline and included in the website in regular intervals. Another promising addition will be to include MNase-seq data [121] as an epigenetic marker gained on the respective cell lines in order to quantify nucleosome occupancy, which has been shown to correlate with cleavage frequency [122].

For the future, we envision to extend the database by appropriately annotated studies targeting high-fidelity Cas9 nuclease variants [37] and nucleases from different organisms [92].

In the following, we want to use this dataset as a training and testing ground for algorithms predicting the cleavage activity of CRISPR/Cas9 for different configurations of cell lines, epigenetic contexts, sgRNAs and more parameters.

Studies

The following studies are included in crisprSQL:

No.	study	assay	cell line	total guides	total targets	epigenetically annotated targets	cleaved gene IDs (CF > 1%, on-targets in bold)
1	Kim	Digenome-seq	HAP1	2	162	7	IGDCC3, KRT42P, CITF22-24E5.1, RP11-638F5.1
2	Cho	targeted PCR	K562	10	116	11	**CCR5, RP11-24F11.2**, CCR2
3	Ran	BLESS	HEK293	4	53	53	**RP11-466M21.1, EMX1**
4	Ran	BLESS	N2A	2	34	0	**TET2**
5	Tsai	Guide-seq	HEK293	4	153	153	KIAA1024, CDH4, CEP89, RP11-115D19.1, RP11-399K21.6,

Figure 2.2: Overview of the included studies, together with metadata such as assay type, the total number of guides and guide-target pairs observed as well as the predominantly cleaved gene names. Studies have been split up according to cell lines.

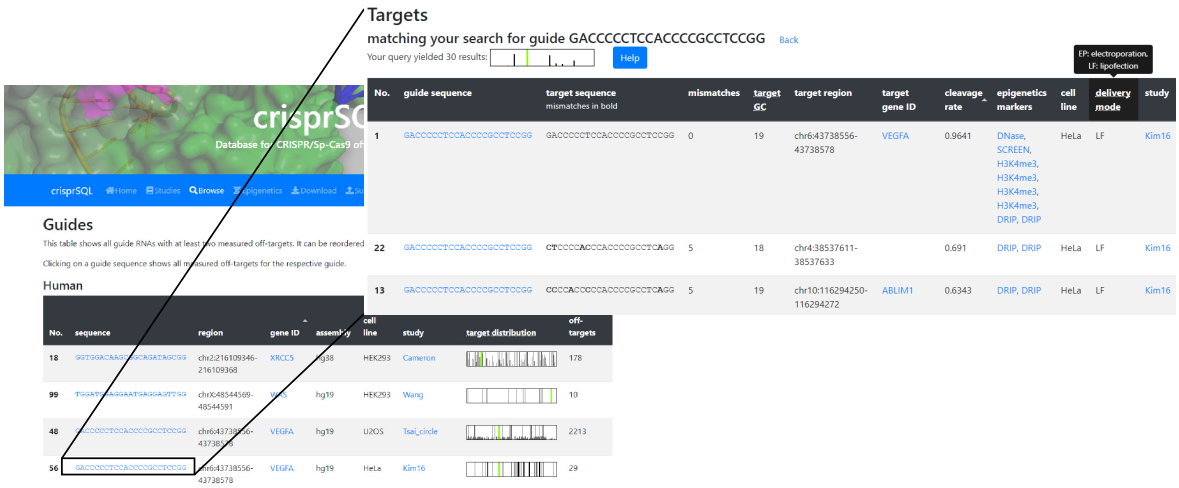


Figure 2.3: Web interface of the database, showing a list of all included gRNA sequences and which, if any, GENCODE gene name they target. Their respective target distributions across the genome are visualised as a barcode-type histogram, allowing a first assessment of their reported specificity. Upon clicking on or searching for a specific guide (inset), the website shows its full reported off-target profile, i.e. gene names of target loci, cleavage rates and epigenetic markers at the target site, hyperlinked to the respective source. Clicking on a gene name opens the vicinity of the respective cut site in a genome browser.

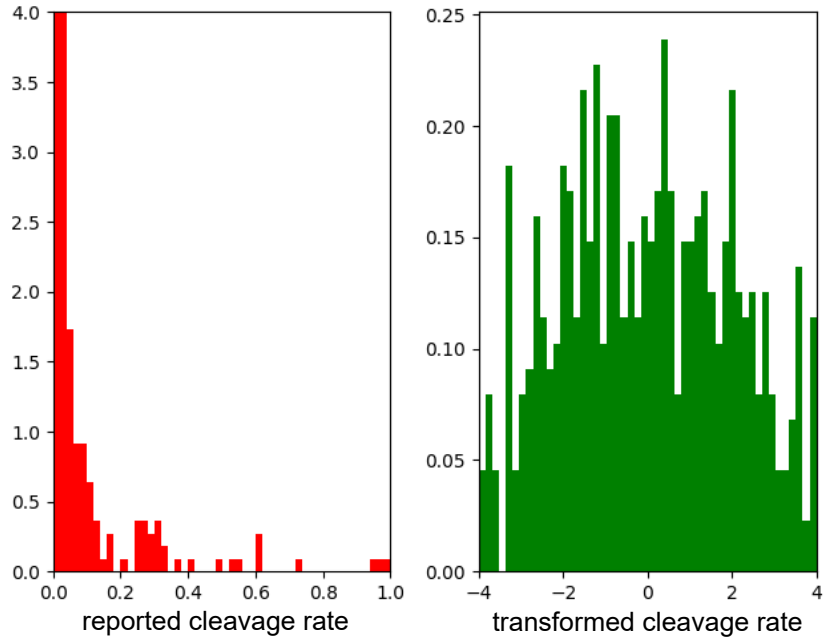


Figure 2.4: Nonlinear transformation of the cleavage rate distribution at the example of study [27]. Red bars (left) indicate the distribution before transformation, green bars after (right). The transformed distribution approaches a Gaussian with mean zero and variance $\sigma^2 = 2$ as desired, with values at ± 4 overemphasised due to capping. A transformation like this is necessary in order to extract a representation of cleavage activity which is free from the biases of experimental studies, e.g. for use in a prediction model.

Chapter 3

Physically informed modelling of CRISPR/Cas9 cleavage efficiency

This chapter has in part been published in *Artificial Intelligence in the Life Sciences* as "piCRISPR: Physically informed deep learning models for CRISPR/Cas9 off-target cleavage prediction" [123]. The generation of nucleosomal features has been conducted in collaboration with a member of our research group and is therefore not detailed below, but the relevant publication is reproduced in Appendix A.1.

Starting from our large, diverse off-target dataset established in the previous chapter, we now turn to implementing novel deep learning algorithms and feature encodings for off-target prediction and systematically sample the resulting model space in order to find optimal models and inform future modelling efforts. We lay emphasis on physically informed features which capture the biological environment of the cleavage site, hence terming our approach piCRISPR. As is conjectured in literature [1], we hypothesise that sequence context and chromatin accessibility will have a considerable effect on cleavage activity and want to test our resulting models against state-of-the-art predictors. We use novel, environmentally sensitive features and scrutinise their relevance for clinical guide design.

This chapter details our journey from first implementing a wide range of models with differing feature sets, feature encodings and model architectures and analysing their performance (section 3.2) to then narrowing down the selection to the two best performing architectures and encodings (section 3.3). These are then subjected to direct comparison with published algorithms.

3.1 Introduction

When considering our dataset, we noticed that repositories of off-target cleavage data contain a significant amount of data points which match in both guide and (off-)target sequence and differ only in the biological environment of the respective loci (see Figure 3.1). Capturing this environment is therefore instrumental in providing accurate predictions of cleavage activity.

In a separate study (see Appendix A.1) we found that computed nucleosome organisation-related features correlate better with cleavage frequency values than experimental epigenetic markers (Deoxyribonuclease-I hypersensitive sites sequencing / DNase, reduced representation bisulfite sequencing / RRBS, CCCTC-binding factor / CTCF, histone-3 lysine-4 trimethylation / H3K4me3) [124] which have heretofore been the literature standard for cleavage prediction models. The computed features also surpassed the aforementioned experimental epigenetic markers in terms of their feature importance in preliminary cleavage activity prediction models which had access to both computed and experimental epigenetic features. We therefore aim to make full use of this novel class of features by embedding them in a rich feature set, including DNA/RNA sequence and context sequence-based features.

3.2 Exploratory modelling

With a considerable amount of cleavage prediction algorithms present in literature [63, 64, 67, 75, 125], we present in this section a model optimisation framework to systematically probe combinations of model architecture, feature set and encoding. We want to provide a one-

stop side-by-side benchmark of a representative set of modelling approaches that take into account novel physically inspired and environmentally sensitive features. Besides improving prediction accuracy and capturing off-target effects that might so far have gone unnoticed, this will also generate insight into the wider biological environment that influences CRISPR cleavage. We start by using a wide range of model architectures, feature sets and encodings. We will then use our learnings to narrow down and refine this in section 3.3, such that we can make confident modelling suggestions whilst retaining a clear narrative.

3.2.1 Data Source

In order to achieve maximum transparency and comparability, we use guide-target pairs from the crisprSQL dataset [83] as detailed in the previous chapter. It is a collection of 17 base-pair resolved off-target cleavage studies on Sp-Cas9, comprising 25,632 data points and is larger than most datasets used to train prediction algorithms to date. It contains data on various cell lines, mainly U2OS, HEK293 and K562. We have chosen to use version 26/05/2020 of the database which does not include T-cell data from [31] in order to avoid introducing a considerable cell line imbalance. Furthermore, the evaluation of our modelling on on-target datasets is beyond the scope of this work due to their different underlying experimental techniques and cleavage quantification measures.

Experimental data points containing guide and target loci, sequence, cell line, assay type and cleavage frequency have been completed and enriched by sequence context as well as five epigenetic markers (CTCF: chromatin organisation [109], DNase: chromatin accessibility, RRBS: DNA methylation, H3K4me3: histone methylation [110], DRIP: R-loop formation in transcription [111]). We combine these established features for cleavage prediction with physically informed features pertaining to the RNA-DNA heteroduplex in order to more faithfully map the CRISPR/Cas9 system. We use empirical binding energy estimates given in crisprSQL based on CRISPRspec [13], where the authors calculate four empirical energy contributions:

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

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

$$\Delta G_B = \delta_{\text{PAM}}(\Delta G_H^{\text{RNA:DNA}} - \Delta G_U^{\text{RNA:RNA}} - \Delta G_O^{\text{DNA:DNA}}). \tag{3.1}$$

For ease of reference, we call this quantity E_{all}. On top of it, we extract the binding energy E_{base} between gRNA and DNA and the free energy released during folding of the sgRNA E_{gRNAfold} to be included in our model (see Figure 3.2) as

$$E_{\text{base}} = \Delta G_H^{\text{RNA:DNA}} = E_4(E_3/E_2), \tag{3.2}$$

$$E_{\text{all}} = \Delta G_B = E_2 - E_{\text{base}}, \tag{3.3}$$

$$E_{\text{gRNAfold}} = \Delta G_U^{\text{RNA:RNA}} = E_1 - E_{\text{base}}, \tag{3.4}$$

where $E_{1..4}$ refer to the energy features from crisprSQL (Table 2.3).

On top of these established features, we propose the usage of nucleosome organisation-related features which add an unprecedented level of sensitivity towards the biological environment of the cleavage site (see Figure 3.1C). In order to reduce the feature dimensionality of the full model and exclude combinations of strongly correlated features, we preselect informative features by training a preliminary cleavage prediction model on 13 distinct nucleosome organisation-related scores. These are all computed based on the 147 bp context around each (off-)target nucleotide (data not shown, see Appendix A.1):

- W/S Scheme: predictor of rotational nucleosome positioning, available on the online database nuMap [126]
- YR Scheme: predictor of translational nucleosome positioning, available on the online database nuMap [126]
- LeNup (H3Q85C) [127]: CNN-based predictor of nucleosome positioning
- NuPoP [128]: duration Hidden Markov Model trained to predict the base-pair specific nucleosome affinity of a given (off-)target sequence. The algorithm yields histone binding affinity, nucleosome occupancy and Viterbi scores.
- nuCpos [129]: Hidden Markov Model-based predictor of nucleosome positioning, building on NuPoP, yields histone binding affinity, nucleosome occupancy and Viterbi scores
- VanDerHeijden [130]: statistical mechanics-based method for predicting nucleosome positioning
- GC147 / GC count: relative proportion of G and C nucleotides within the 147 base pair window around a given target nucleotide
- Nucleotide BDM: training-free method that approximates the algorithmic complexity of a DNA sequence by applying the block decomposition method [131] on the 147 bp sequence around each nucleotide. Low values of Nucleotide BDM have been shown to correlate with proximity to nucleosome dyad positions [131].
- Strong-Weak BDM: Nucleotide BDM starting from the sequence which has G/C replaced with S and A/T replaced with W

Besides these features, the model also had access to the four literature-standard epigenetic markers CTCF, DNase, RRBS, H3K4me3 as defined above. The three scores of the highest overall feature importance value for both a tree-based model and a CNN were GC count, Nucleotide BDM [131] and NuPoP Affinity [128]. For the parameters which were used to

calculate these three features, we refer to [124]. We want to emphasise that since these features are computed purely from the sequence context, they are not subject to effects from chromatin state or cell cycle, in contrast to epigenetic features gained experimentally from cell lines.

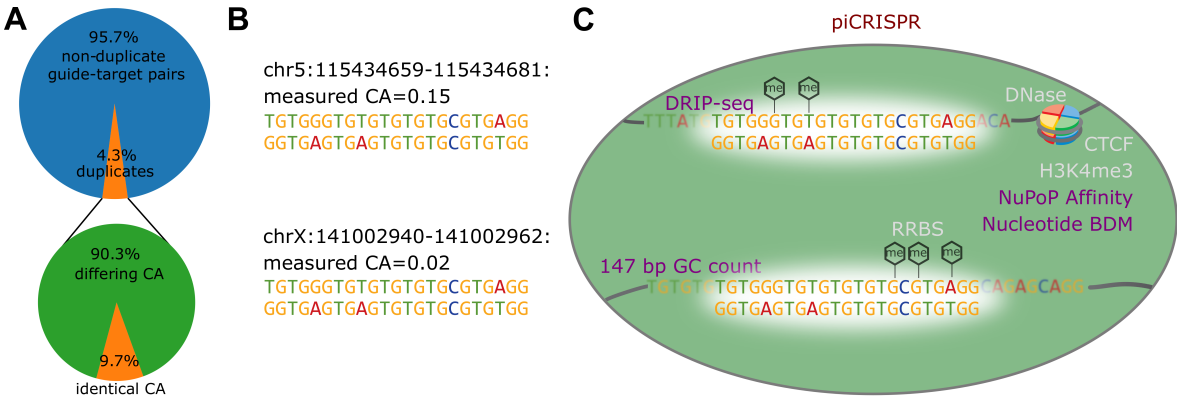


Figure 3.1: **A** The crisprSQL dataset contains an appreciable amount of perfect guide-target duplicates. We only consider data gained from human cell lines and putative off-targets which we generated based on sequence similarity. 8,922 of 230,274 data points have at least one guide-target duplicate within this set which differs in cleavage activity (CA). As the example from our dataset in panel **B** shows, such a pair looks identical to purely sequence-based prediction algorithms. They might therefore not predict dangerous off-target effects. **C** piCRISPR remedies this by taking into account the biological environment of the cleavage site based on a range of features beyond guide and (off-)target sequence. So far, prediction algorithms have used features related to chromatin organisation (CTCF, [109]), chromatin accessibility (DNase-Seq), DNA methylation (RRBS) and histone methylation (H3K4me3, [110]). We further use features pertaining to R-loop formation during transcription (DRIP, [111]) and the following scores based on the 147 bp sequence context around each (off-)target nucleotide: GC count, sequence complexity (BDM, [131]) and nucleosome positioning information (NuPoP, [128]) which introduce unprecedented sensitivity to the biological environment of the cleavage site. Using these, piCRISPR can correctly rank the two example loci given here.

Labels

We use the cleavage activity (CA) values given by crisprSQL as basis for the labels. These refer to the relative frequency with which a given cleavage event has been observed and do not take into account subsequent repair of the cleavage site. That process might be described by concepts such as repair efficacy, i.e. the relative frequency with which the cleavage site is repaired with a certain outcome, and mutagenic repair rate, i.e. the kinetic rate constant of imperfect cleavage repair. For classification, we define the negative class as all data points with cleavage activity values below the lowest reported assay accuracy of 10^{-5}, combined with the set of putative off-targets.

In order to achieve comparability between different studies for regression tasks, we perform a nonlinear Box–Cox transformation [120] to transform the cleavage rates to approximate a Gaussian with zero mean and variance $\sigma^2 = 2$, similar to the approach in [60] and [119]. The nonlinearity in this approach also means that the regression label is less sensitive to measurement noise for data points with low cleavage activity. Cleavage activity values below the lowest reported assay accuracy of 10^{-5} as well as putative off-targets were set to $-2\sigma^2 = -4$, and transformed values subsequently clipped to the $[-4, 4]$ range. This is an empirical choice based on the shape of the resulting distributions.

3.2.2 Data Augmentation

In order to increase the size of the training set, we extend it by those putative off-target sites along the respective genome which had fewer than seven mismatches to each respective guide sequence, omitting the (off-)target locus itself. Putative off-target sites were found using the k-mismatch sequence aligner `batmis` [132] which promises speedup compared to tools like `bowtie` and the Burrows–Wheeler aligner `BWA`. The following commands were used for this:

```
batman -q filename -g genomePath -n nMismatches -mall -o hits_experiment_id_genome.txt
batdecode -g genomePath -i hits_experiment_id_genome.txt -o out_experiment_id_genome.txt
```

Another option is to use the Cas-OFFinder [73] which has the option of aligning with RNA bulges, as was done in the case of CRISPR-Net [67]. The following commands can be used for this:

```
echo '' >> filename | sed '/^>/d' filename | sed ':a;N;$!ba;s/...\ n/NNN
nMismatches\ n/g' | sed  '1s;^;genome_filename\ nNNNNNNNNNNNNNNNNNNNNRG\
n;' > filename2
cas-offinder filename2 C outexperiment_id_genome.txt

echo '' >> filename | sed '/^>/d' filename | sed ':a;N;$!ba;s/...\ n/NNN
nMismatchesBulge\ n/g' | sed  '1s;^;genome_filename\ nNNNNNNNNNNNNNNNNNNNNRG
1 1\ n;' > filename3
cas-offinder filename3 C outexperiment_id_genome_bulges.txt
```

After retrieving the alignments, it was ensured that the protospacer adjacent motif (PAM) was either the canonical 5′-NGG-3′ characteristic of SpCas9 [22], or one of the noncanonical forms 5′-NGA-3′ and 5′-NAG-3′ observed in [21]. If a genome-wide off-target detection method has not detected cleavage at a locus within the genome that satisfies these criteria, we deem the cleavage activity at this point to be zero. This yielded 310,142 total guide-target pairs, making the complete data set highly imbalanced. Sticking with the convention in literature, we refer to this process of extending the number of data points as *data augmentation*.

For this work, we concentrated on the 251,854 data points which originated from a human cell line or synthetic human DNA.

3.2.3 Feature Encoding

In order to be able to train a variety of models, we first encode the features as a single vector per data point (Figure 3.2A). We call this the 'target-guide encoding' (E0) since target and guide are represented separately. Sequence information of both guide and target are concatenated as consecutive one-hot vectors. This vector is further concatenated with a vector of epigenetic markers (as was done in [75]) which are normalised to $[0, 1]$ for individual markers. We further concatenate this with CRISPRspec energy and nucleosomal features. The former were normalised uniformly such that their overall maxima and minima also are in $[0, 1]$. Nucleosomal features were taken from the respective algorithms without additional normalisation. Exploring latent representations of guide or target is not within the scope of this work, given that it further complicates comparison between models.

Based on the energy-driven nature of binding and cleavage, we hypothesise that mismatched interfaces affect binding in a totally different way than matched interfaces. This has so far not been recognised in detail by off-target prediction algorithms. In order to explicitly include the information at which a given target-guide pair contains a mismatched interface, we introduce the 'target-mismatch' encoding (E1, see Figure 3.3). Here, we split the target sequence into matched and mismatched interfaces, encoding the matched interfaces in a one-hot fashion first and then append the one-hot encoding of the mismatched interfaces. This encoding loses information about the precise type, i.e. the base pair identity of the mismatch.

In order to remedy this, we introduce the 'target-mismatch type' encoding (E2, see Figure 3.2B), which also encodes the type of a respective mismatch. It gives the most detailed insight into the nature of a given base-pair interface.

The 'target-OR-guide' encoding (E3, see Figure 3.2C) combines the one-hot encoded versions of guide and target using an entry-wise OR operation between the vectors, as first seen in [66].

For convolutional models, we reshape the resulting vector into a matrix, in which the 23

base pairs each have separate channels for sequence, epigenetic, nucleosomal and energy information (Figure 3.2D). Where a model requires separate inputs for guide and target, we perform this operation twice with only the respective sequence contributing. This type of matrix encoding is inspired by Chuai *et al.* [75].

3.2.4 Model Architectures

Literature contains a wealth of model architectures commonly used to predict CRISPR cleavage. Currently, successful model architectures for learning-based cleavage prediction fall in one of three categories [133]: tree-based methods [60], convolutional neural networks (CNN, [63, 75]) and recurrent neural networks (RNN, [67, 108]). Thus far, a rigorous comparison between these is missing in literature, and some have only been applied to on-target activity prediction [108]. We therefore take successful CNN and RNN architectures present in the field and adapt them to the task of off-target prediction using various encodings of the subset of features described above.

Our CNN model is based on the architecture described in [75] (see Figure 3.2E). There, the outputs of two separate, convolutional layer-based encoders for guide and (off-)target are concatenated channel-wise (forming the Siamese part of the network) and serve as input for a convolutional classifier (the conjoined part of the network). Instead of a dropout layer, we use a Gaussian noise layer after the first convolutional layer in each encoder since we found this to improve validation set prediction benchmarks. Having batch normalisation layers only before and after the Siamese layers increased training stability without compromising training performance. The same was true for applying (leaky) ReLU activation functions after the conjoined (Siamese) layers. The former also contained a dimension-preserving max-pooling layer in between the convolution and ReLU layers. Furthermore, we found that in our case, removing the 256-filters convolutional layer before concatenation of the Siamese channels did not considerably affect training performance but reduced training time. We then adjusted the downstream network accordingly, which involved removing the penulti-

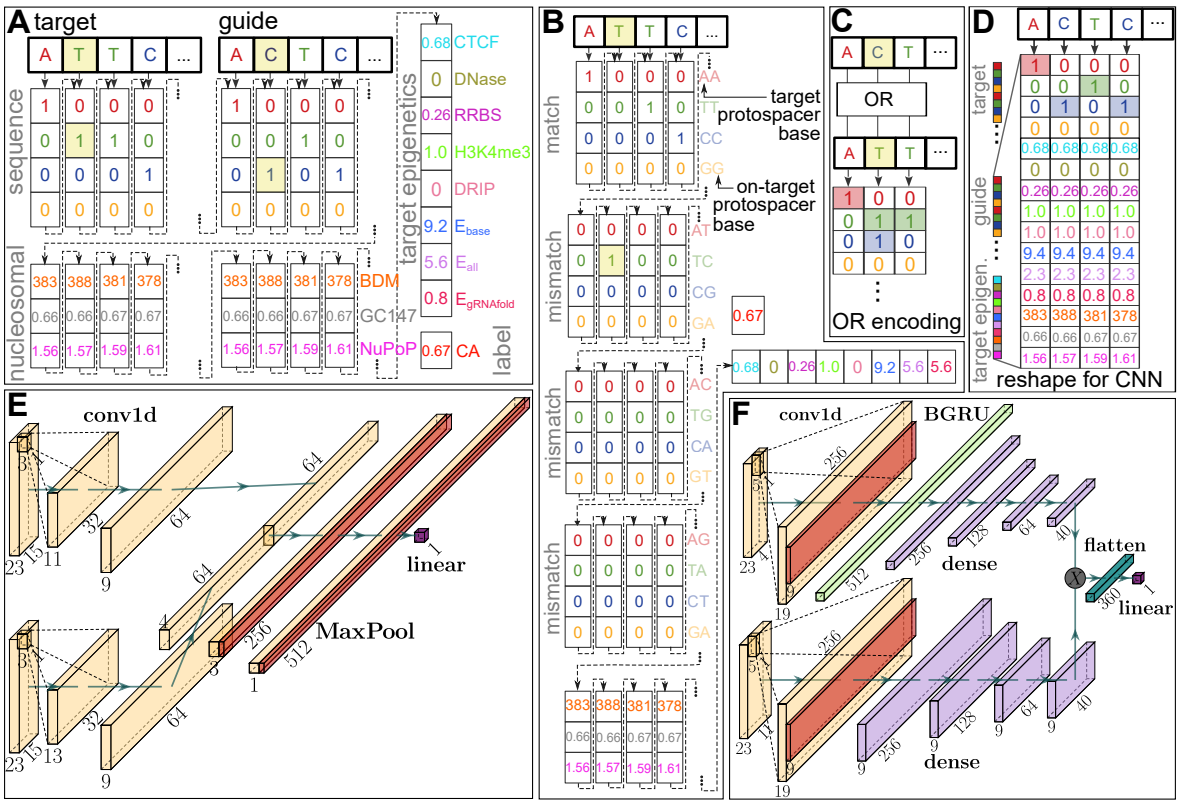


Figure 3.2: Overview of feature encodings. **A** Target-guide encoding (E0): (Off-)target and guide (on-target protospacer) sequences are one-hot encoded and concatenated together with the epigenetic features (CTCF: chromatin organisation, DNase: chromatin accessibility, RRBS: DNA methylation, H3K4me3: histone methylation, DRIP: R-loop formation during transcription, E: free energy estimates from CRISPRspec [13, 118]) and nucleosome positioning channels (147 bp GC count, Nucleotide BDM [131], NuPoP Affinity [128]) of the target to form a single column vector per data point. The Box–Cox transformed cleavage activity (CA) is used as a label. In target-mismatch encoding (E1, see Figure 3.3), the four-letter alphabet is doubled, with the first part encoding the target and the second part indicating whether the base pair on the target strand is canonically matched to the guide, i.e. identical with the protospacer base pair. **B** Target-mismatch type encoding (E2) extends this encoding using a 16-letter alphabet which also contains information about the precise nature of the mismatch. **C** Target-OR-guide encoding (E3) is an efficient way of representing both target and guide sequence in a single four-vector per nucleotide by performing an OR operation between the one-hot encoded target and guide nucleotide vectors (as used in [66]). **D** In order to serve as input for a convolutional neural network, the vector is reshaped into a $(4m + n) \times 23$ matrix where n is the number of non-sequence channels and $m = 1$ for E0 and E3, $m = 2$ for E1 and $m = 4$ for E2. **E** Siamese convolutional neural network architecture used in our CNN model, based on the model in [75]. For encoding E0, target and guide sequence are encoded as separate matrices as shown in panel C and serve as the inputs (left side). Target epigenetics and nucleosomal features are copied to the corresponding channels on the guide arm. For encodings E1–E3, only one arm of the Siamese portion of the network is used, with its number of input channels set to $4m + n$. The channel dimensions given in this Figure are valid for the S5 feature set, see Table 3.1. **F** Bidirectional gated recurrent unit architecture as used in our RNN model, based on the model in [108]. Sequence features (upper branch) are encoded in E3 encoding; epigenetics-based features (lower branch) are encoded as shown in panel D.

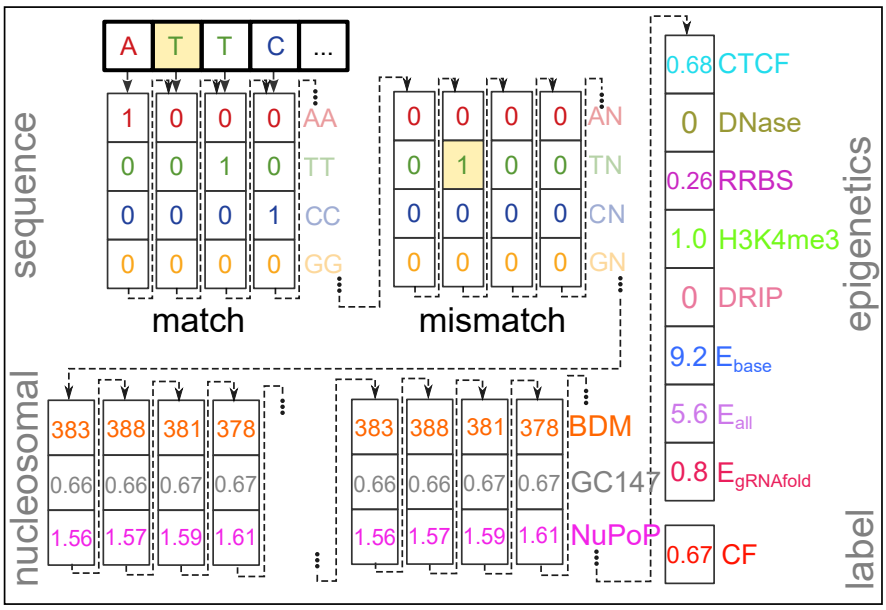


Figure 3.3: Target-mismatch encoding (E1) as described in Figure 3.2.

mate convolutional layer and halving the number of filters in the remaining convolutional layers. For classification, we choose a sigmoid rather than linear activation function for the last layer. The result of this is shown in Figure 3.2E.

The RNN model is based on the bidirectional gated recurrent unit (BGRU) on-target prediction model from [108]. Here, a BGRU layer is used to make use of the relevant longer-range dependencies between sequence features that would go unnoticed by a CNN of manageable kernel size. In order to make this type of architecture usable for off-target prediction, we combine guide and target sequence using an element-wise OR operation as seen in [66] (see Figure 3.2C). The size of the model's epigenetics channel was increased to accommodate epigenetics, energy and nucleosomal features in a matrix encoding (see Figure 3.2F).

3.2.5 Model Training & Evaluation

Given the imbalance of validated/measured and non-validated/augmented data points, we employ a bootstrapping strategy as suggested in [86], where training batches on average contain equal numbers of both classes. For regression (classification), early stopping is based on the mean squared error (binary cross-entropy) loss on half of the test set, where the other half is reserved for evaluation.

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

Table 3.1: Abbreviations for feature sets S0-S5 used in this section.

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

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

The RNN models are trained for 100 epochs, where batches of 10,000 points are sampled each epoch out of a class-balanced sample of 50,000 data points. Replicating the transfer learning approach taken in [108], we then freeze the network up to and including the second fully connected layer on both sequence and epigenetics branch and train the remaining layers for a further 100 epochs. For testing scenario 1, both training runs are subject to early stopping with a patience of 4 epochs. For testing scenario 2, early stopping was not used but instead a fixed number of 120/25 epochs before/after freezing, respectively. Dropout probability was set to 0.2 and the Adam learning rate was at $1 \cdot 10^{-3}$.

Exploring the model space

In order to systematically compare the training performance of combinations of (regression/classification, model architecture, feature selection, feature encoding), we performed individual training runs for all possible combinations within this landscape. To make this independent of our choice of data augmentation, we only included experimentally validated data points within the held out 20% of the crisprSQL dataset in the testing set. The results of this can be seen in Figure 3.4. We then took only the six best-performing models in terms

of both Spearman r (for regression) and area under receiver operating curve (AU-ROC, for classification) further into other testing scenarios detailed below. The six best-performing models were spread across all three architecture types, with CNNs yielding the best performance benchmarks and thus providing three of the best-performing configurations. RNNs and gradient-boosted tree models contributed two and one configurations, respectively.

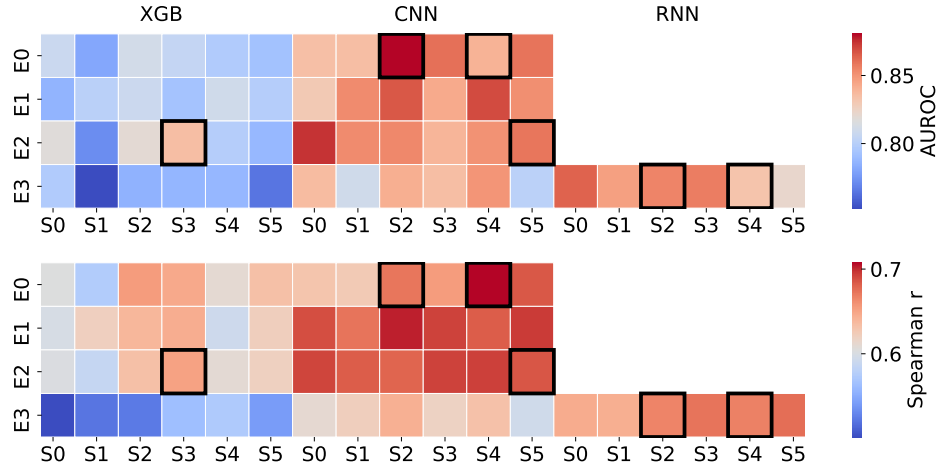


Figure 3.4: Validation benchmark values for various combinations of feature sets (S0-5, see right panel), encodings (E0-3, see Figure 3.2) and models (gradient-boosted tree/XGB, convolutional/CNN, recurrent/RNN). Models were trained on a randomly sampled 80% of the crisprSQL dataset [83] and tested on the validated points within the remaining 20%. Rectangles indicate the combinations which we have chosen to scrutinise further.

Testing Scenario 1: crisprSQL comparison

This testing scenario uses the whole crisprSQL dataset in an 80/20 split between training and testing. As suggested in [86], we use all validated and non-validated points within our dataset. Augmenting the measured data points with putative off-targets leads to a class imbalance of 1:10.85 (measured:augmented).

In order to assess the practicality of a large multi-study dataset, we obtained an estimate of the saturation of the data absorption for certain models with respect to the number of studies in training and test data. We gradually introduced training and testing data in ascending order of training performance, i.e. the study with the highest prediction benchmark

when training/testing on its 80/20 split being introduced first. The results of this can be seen in Figure 3.6.

Scenario 2: literature comparison

In this scenario we hold out studies [90, 96, 97] from the training set. These studies have not been included in the training set for the state-of-the-art off-target prediction algorithm CRISPR-Net [67], such that they remain an independent test set to compare CRISPR-Net and piCRISPR side by side. This scenario is our default way of evaluating piCRISPR performance (see Figures 3.9 and 3.11). The inherent class imbalance in this test set is 1:103.96.

Scenario 3: set of duplicate pairs

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

Pairwise training When combining various experimental studies, care must be taken as to their interaction during training. We therefore devised the notion of a pairwise training performance of two studies, which we define as the training performance when training on the larger and testing on the smaller of these studies. We did this for all studies in the

crisprSQL dataset and visualised the result in a force-directed graph (Figure 3.7, Figure B.1) where the force between nodes is proportional to the third power of the Spearman r obtained by pairwise training and testing.

Model Explanation

Using the model-agnostic Shapley Additive Explanations (SHAP) library [134], we obtain feature importances for each of the input features. Since piCRISPR wraps the feature encoding inside a given model, we retain full explainability of input features even for non-invertible encodings such as E3. Using interface encodings E1 and E2, we obtain an unprecedented, context-sensitive resolution of sequence-based features.

Sticking with the convention set by the SHAP library, we calculate global SHAP values as the mean of the absolute value of SHAP values across data points in the explanation set, which is a random subset of 500 points from the held out test set. In order to show not only the magnitude but the direction in which a given feature influences the model's prediction, we multiply each feature's global SHAP value by the sign of the average SHAP value of all data points whose value is larger than the median of that feature.

Command line usage of our models

We have implemented a command line interface with which piCRISPR predictions can conveniently be obtained. In order to achieve maximum usability, the model automatically uses default feature values in case a certain feature was not provided, thereby lowering prediction performance (see Figure 3.5). The default value of a given feature is defined as the average feature value of the set of those crisprSQL data points which lie within a certain box around the mean SHAP value. This means that high-accuracy piCRISPR predictions can be obtained in a user-friendly way, even when providing only guide and (off-)target sequence. Our online repository contains hands-on examples on this.

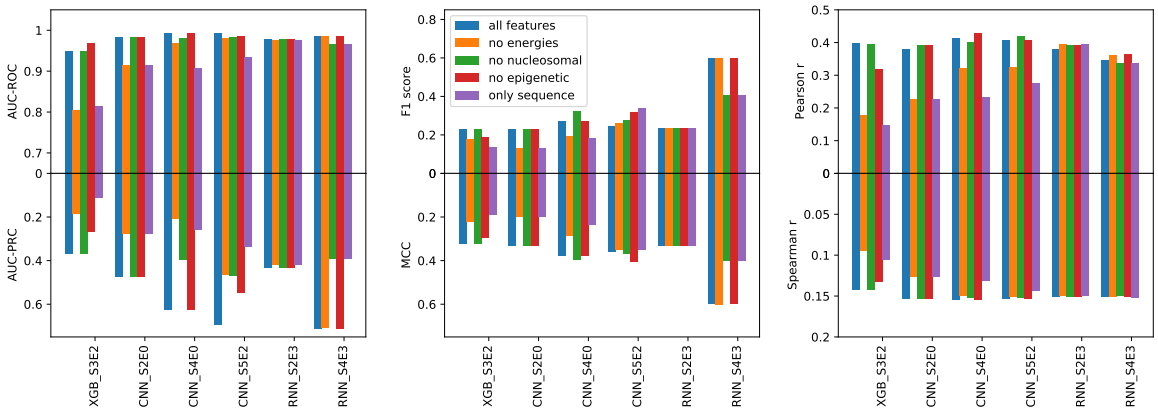


Figure 3.5: Performance comparison between piCRISPR models when tested on held out studies [90, 96, 97] whilst ommitting certain sets of features. piCRISPR then uses default values for these features, leading to slightly reduced prediction accuracy.

3.2.6 Results & Discussion

Testing Scenario 1

Data portioning Figure 3.6 shows that when gradually increasing training data study by study, with the highest-performing study introduced first, CNN models appear to generalise fastest, whereas RNN models appear to absorb data more gradually. We also find that all models absorb data more gradually when measured by AUC-PRC than by AUC-ROC or Spearman r, which supports the notion in [86] that AUC-PRC is the more relevant benchmark parameter here.

Pairwise training Figure 3.7 shows that even though studies using the same cell lines show some bunching in a force-directed 2D graph where distance is proportional to generalisation from one study to the other, bunches heavily overlap between cell lines and experimental conditions. This supports the mixing of training data across studies despite differing experimental parameters. Experimental studies conducted on HEK293 cells appear to generalise best. Three distinct studies appear to generalise comparatively poorly to other studies, one of which was gained on HAP1 cells. Study [99] which has been gained on synthetic DNA appears to generalise to other studies about as well as studies on K562 cells.

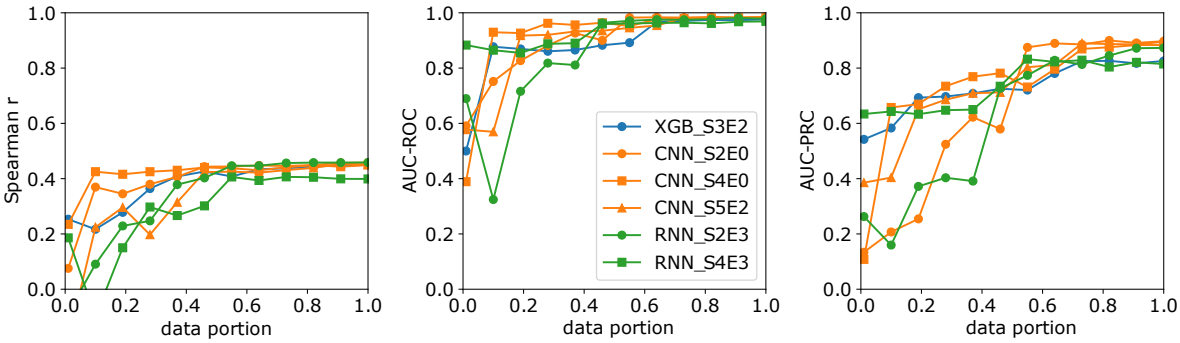


Figure 3.6: Training performance of the six models when gradually introducing training data study by study along the x-axis, ordered by single-study performance with the highest-performing study introduced first (at low data portion values). Models were trained on a randomly sampled 80% of the crisprSQL dataset [83] and tested on the remaining 20% (testing scenario 1), meaning that a maximum of 201,483 data points were used in the training set. Plot points are equally spaced in terms of data portion; solid lines are shown as a guide to the eye. We can see that CNN-based models (orange) seem to generalise fastest, i.e. for lowest data portion values, whereas RNN models appear to absorb data more gradually. Models appear to saturate around the 60% mark or 120,000 data points, indicating that the weakest-performing 40% of studies might be used for purposes other than training the classifier/regressor.

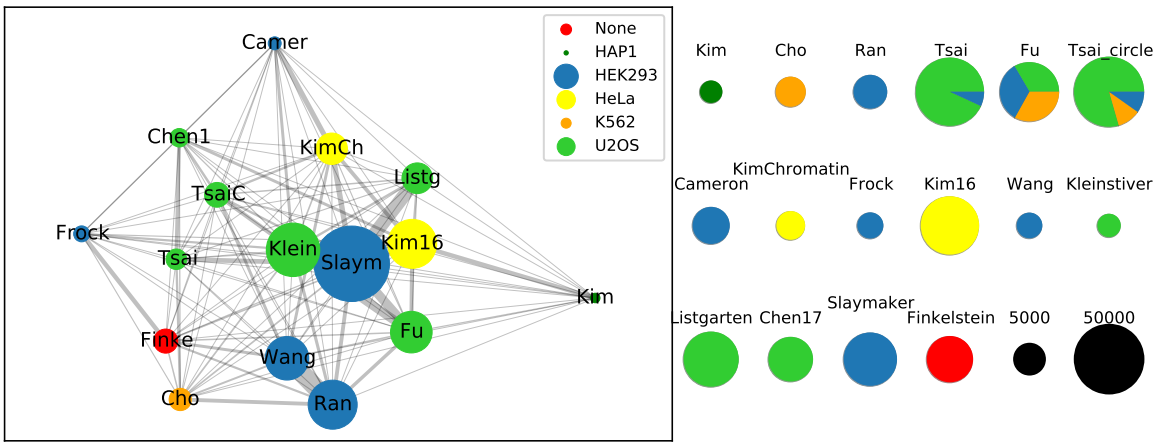


Figure 3.7: Left panel: 2D distance network representation of pairwise training performance using a Fruchterman-Reingold force-directed algorithm. The force between two nodes (studies) is proportional to the third power of the Spearman correlation when training on the larger and testing on the smaller of these studies. Study labels are abbreviated for better visibility. Edge width represents pairwise training performance of the two adjacent studies alone; bubble size indicates 'study importance', i.e. the overall summed performance of a study. Close positioning indicates good pairwise training performance. Studies have been coloured by majority cell line; all data has been gained using the CNN_S4E0 model which was shown to generalise fastest in terms of Spearman r in Figure 3.6. **Right panel**: Composition of the cell lines making up each individual study (colours as in left panel) with sizes proportional to the number of data points per study, including non-validated data points. Black circles act as a size legend.

BDM cutoff value During the process of analysing our dataset, we noticed that the StrongWeak BDM feature shows a distinctly different distribution when looking at data points in the lowest cleavage activity class CA $= -4$ after Box–Cox transformation. Figure 3.8 shows this and sparked our interest in a closer quantitative analysis of how epigenetic descriptors affect CRISPR-Cas9 off-target activity, ultimately resulting in our publication [124]. This could be due to the demonstrated correlation of low BDM values for a certain DNA sequence with a high likelihood for a nucleosome dyad to be positioned there [131], which results in reduced binding.[1] Since both binding and cleavage are required for a data point to be labelled as active in our data, we hypothesise that excluding training data points with a low average StrongWeak BDM value reduces the number of putative training data points for which binding did not take place, and thereby lead to the model picking up better on features pertaining to cleavage. We show the generalisation performance of this in the bottom panel of Figure 3.8, where we can see a slight improvement in AU-PRC for a StrongWeak BDM cutoff value of 320 compared to the unfiltered dataset (cutoff = 0). For other testing scenarios, the improvement was even smaller, such that we have chosen to not apply this filtering in the following subchapter. For further reading on this topic, we instead refer to our publication [124] which is reproduced in section A.1.

Testing Scenario 2

Figure 3.9 shows the regression and classification performance of our piCRISPR-implemented models, with CNN_S5E2 yielding the highest benchmarks. As mentioned in [86], the area under precision-recall curve (AU-PRC) is a much more suitable measure than the area under receiver operating curve (AU-ROC) for off-target prediction, since in clinical application,

[1] Another possible explanation would be the presence of polymerase-III terminating runs of A nucleotides on the sgRNA which reduce sequence complexity and would suppress its expression. However, at least four consecutive As are needed to induce termination [135], which is only given in less than 0.5 % of the sgRNAs in our dataset.

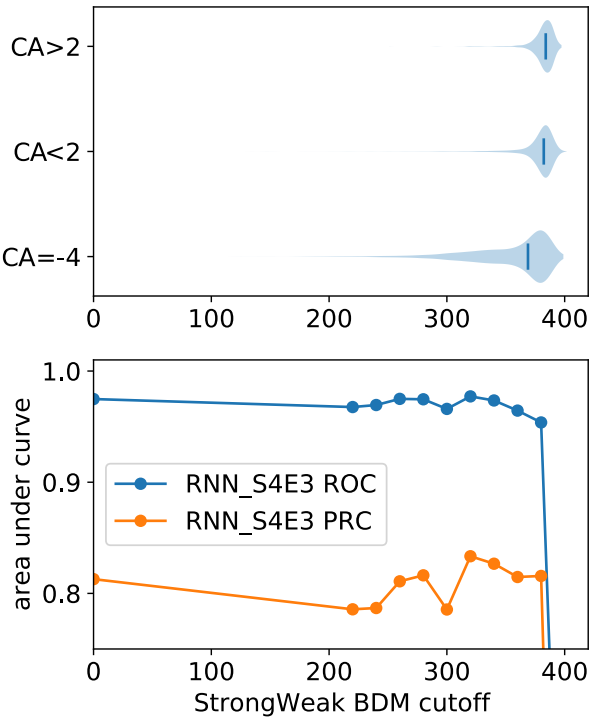


Figure 3.8: Top panel: Violin density plots of the StrongWeak BDM feature, individually for three value ranges of the Box–Cox transformed cleavage frequency. The distribution of StrongWeak BDM values appears to have a longer tail and lower mean value (indicated by a vertical bar) for the lowest cleavage activity class (CA $= -4$). This observation sparked our publication on epigenetic descriptors affecting CRISPR-Cas9 off-target activity [124]. **Bottom panel:** Area under ROC and PRC curves for the RNN_S4E3 model in testing scenario 1 (as in Figure 3.6), where data points with an average StrongWeak BDM value below a certain cutoff have been removed from the training set.

false negatives have far more adverse effects than false positives.

When considering AUC-PRC as a measure, the S4 and S5 feature sets, which include nucleosome positioning-related features, outperform the literature-standard sets S2 and S3 which do not (see Table 3.1). This supports our hypothesis that nucleosomal features are a key ingredient to cleavage prediction.

A direct comparison with prediction results obtained from CRISPR-Net on an identical test set which neither algorithm has been trained on can be found in Figure 3.11. piCRISPR achieves higher prediction benchmarks for both classification and regression, except for the Matthews Correlation and F_1 score on study [90]. We attribute this to the large cell line di-

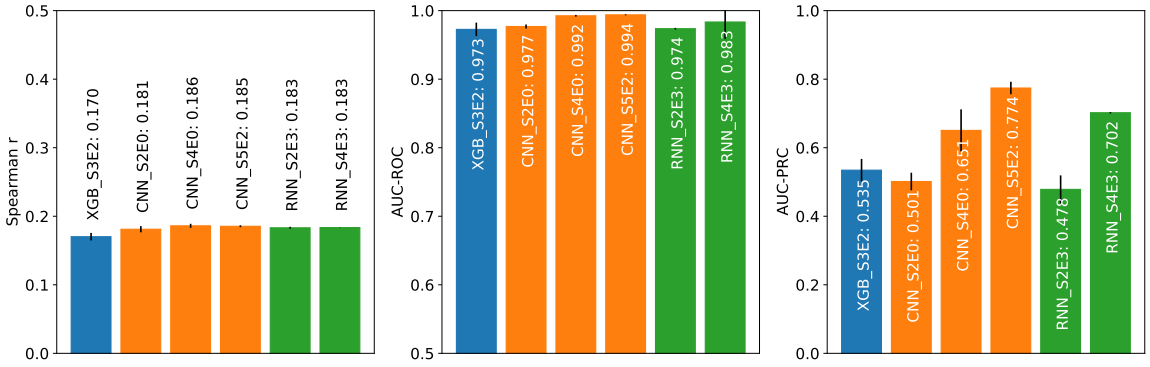


Figure 3.9: Comparison of models, using various methods, features and encodings. The models were tested on the held out studies [90, 96, 97] and trained on the remaining studies within the crisprSQL dataset [83] (testing scenario 2). Non-validated data points have been oversampled in the test set to match the class imbalance of 1:79.35 found in the dataset I-1 from [67]. All models have been implemented in piCRISPR, where training and testing have been repeated 5 times to obtain mean and standard deviation as shown. For the underlying ROC and PRC curves see Figure 3.10.

versity in this study. We note that we have observed an inverse dependence of Spearman correlation on class imbalance ratio (data not shown), which necessitates an identical class imbalance ratio in the respective test sets. Our CNN_S5E2 model consistently outperforms CRISPR-Net on the identical held out test set both in terms of Pearson and Spearman correlation.

Feature importance Figure 3.12 shows that the CNN_S5E2 model draws on sequence features which stem from mismatched interfaces differently than on those from matched interfaces, supporting our hypothesis that this differentiation is not only physically indicated but also backed by the model's behaviour. Global SHAP values suggest that the preference of the variable PAM nucleotide at position 21 is contingent on the specific sgRNA–DNA interface formed. Whilst we recover the preference for G as well as the disfavouring of T found in literature [75, 136] for matched interfaces, it appears different for mismatched interfaces. The model extracts G to be favoured when mismatched with G on the guide, to be disfavoured when mismatched with A on the guide. This supports the notion that a concentration on

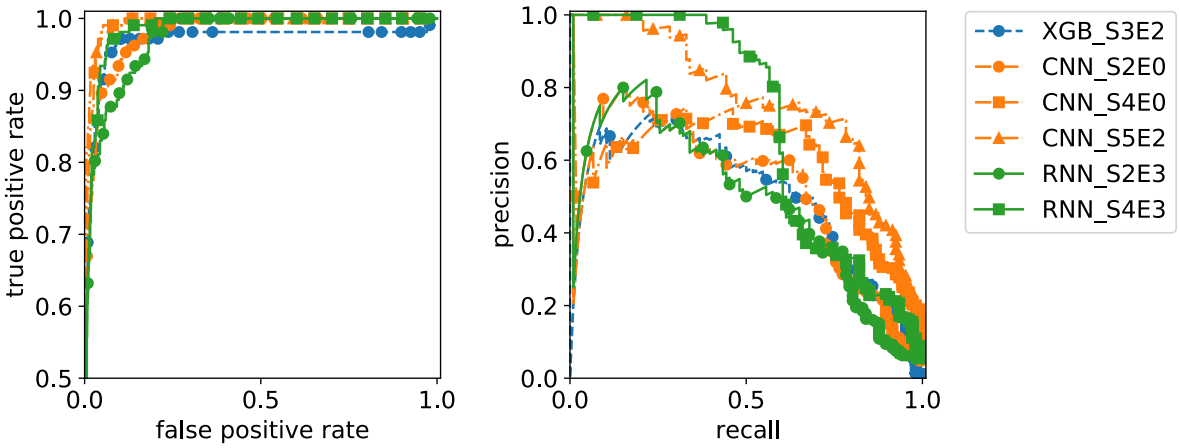


Figure 3.10: Receiver operating (left) and precision-recall (right) curves underlying Figure 3.9.

guide-target interfaces rather than pure base identities is necessary for off-target prediction, and that deeper insight is required than the notion of a preferred base at a specific position. The same goes the notion of a favoured guanine / disfavoured cytosine at position 20 of the sgRNA, which we recover only for matched interfaces. These observations all suggests that it is necessary to consider mismatch interfaces with sequences in the desired genome, not just the sequence of the putative guide, for sgRNA design. We cannot recover the notion of a PAM-distal 'preference zone' and a PAM-proximal 'avoiding zone' of mismatches as seen in [75] when averaging over feature importance values pertaining to each nucleotide, which might be due to our considerably more diverse training dataset. However, we do observe a mismatch-induced reduction in cleavage activity at nucleotides 3–5 and 9–11, in accordance with a recent cryo-EM structural study [137] that has shown that mismatch detection by the REC3 domain of Cas9 is high in these regions. Note that due to the low prevalence of non-NGG PAMs in our dataset, as has been our choice when augmenting it with putative off-targets, the model attributes little importance to the two 5′ GG base pairs.

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

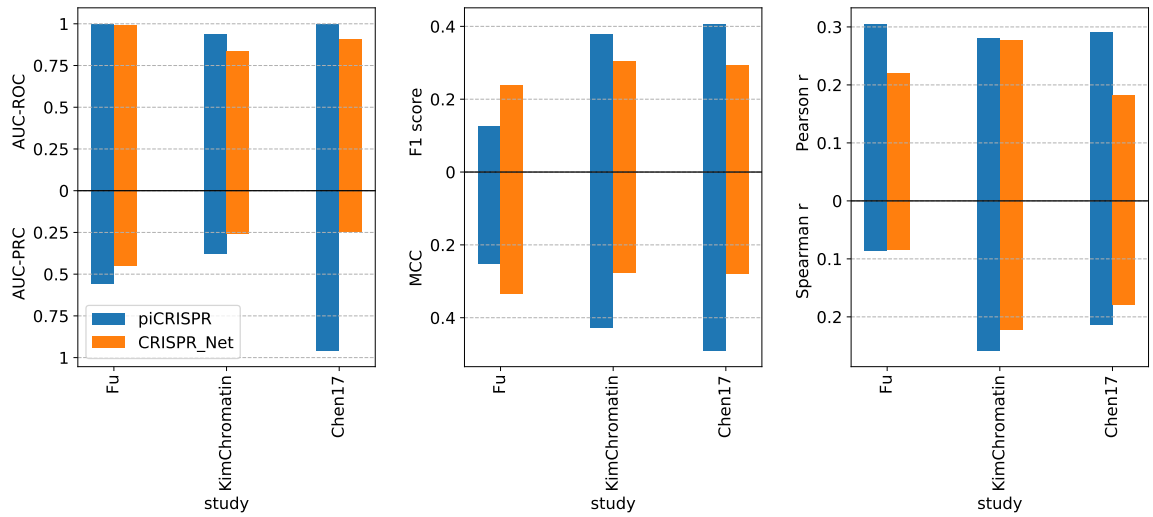


Figure 3.11: Comparison between piCRISPR (model CNN_S5E2) and CRISPR-Net [67]. Both models were tested on the identical set of studies [90, 96, 97] (testing scenario 2) which neither model has been trained on. Aside from the areas under receiver operating and precision-recall curve, we also evaluate the Matthews correlation coefficient (MCC) and the F_1 score as classification benchmarks. Predicted scores were gained using the command line feature available for both models.

off-target cleavage [136]. We further observe that the NuPoP Affinity score ranks higher in terms of global SHAP value than most sequence features and all established epigenetic features. The negative influence of nucleosome affinity can be explained by the reduced accessibility of high-affinity DNA regions, and is observed strongly between nucleotides 5 and 19.

This also demonstrates the importance of nucleosome-related features for cleavage prediction, and also supports the notion of chromatin accessibility influencing cleavage activity found in [138]. To our knowledge, this strong effect of a more than 10 bp wide sequence context on genome-wide off-target cleavage prediction has not been demonstrated yet. Hints of it have been seen only for smaller contexts and on-target efficacy prediction [139, 140]. In addition, our findings present an unprecedented example in which information in the 147 bp sequence context has considerable impact on the model.

Testing Scenario 3

Table 3.2 shows that the model performance, measured by the mean squared error of predictions within a group of data points that share both guide and target sequence, is considerably decreased by introducing features beyond sequence information (left column). We find that the lowest error value is achieved when setting the literature-standard epigenetic scores first introduced in [75] to a default value across data points, indicating that they might not interact favourably with other features and not aid our model's absolute predictions. The resulting distribution of MSEs is shown in Figure 3.13. Setting nucleosomal features to a default value leads to an increase in error value, supporting our notion that these contribute to model accuracy.

Looking at the relative pairwise difference, we observe that the introduction of features beyond sequence leads to an increase of the average proportion of true cleavage frequency differences between points of differing biological environment which is captured by the model. Whilst the full feature set (see Table 3.1) achieves the highest proportion, setting the epigenetic scores to their default value only has a limited impact on this value, further indicating that their encoding of the biological environment of the cleavage site is of limited importance to the prediction. In contrast, the proportion drops considerably when setting nucleosomal values to their defaults. This supports the notion that the biological environment of the cut site is captured more accurately by nucleosome positioning-related features (highlighted in Figure 3.1) than by the literature-standard epigenetic scores (greyed out there).

3.2.7 Conclusion

The key learnings from this exploratory study regarding the choice of model architecture, feature sets and encodings are as follows:

- Both CNN and RNN architectures deliver a literature-comparable performance. The

tree-based XGBoost model, while computationally efficient to train and retrieve feature importance values, shows the lowest performance benchmarks. This is likely due to the absence of a notion of nucleotide locality which is explicit in the data and model structure for CNN and RNN models. The tree-based model will therefore not be considered any more.

- Since only two of the three energy features show high global feature importance values, we chose to use the CRISPRoff score [13] as a weighted combination of the separate contributions we list here.

- The four traditionally used epigenetic features only have limited impact. We therefore do not include them any more in the next section.

- The heavily skewed cleavage activity distributions complicate the calculation of meaningful benchmarks for the regression task. This is further exacerbated when including putative off-targets. The regression approach is therefore only appropriate for a well-defined class imbalance in the test set, which is generally not the case in literature.

- The target-mismatch type encoding (E2) offers feature importance values of unprecedented resolution.

- The target-OR-guide encoding (E3) works well for the RNN architecture. It avoids an all too sparse encoding at the expense of information about the directionality of the mismatch being list. We therefore want to add another dimension to alleviate this loss of information and apply it to the CNN as well.

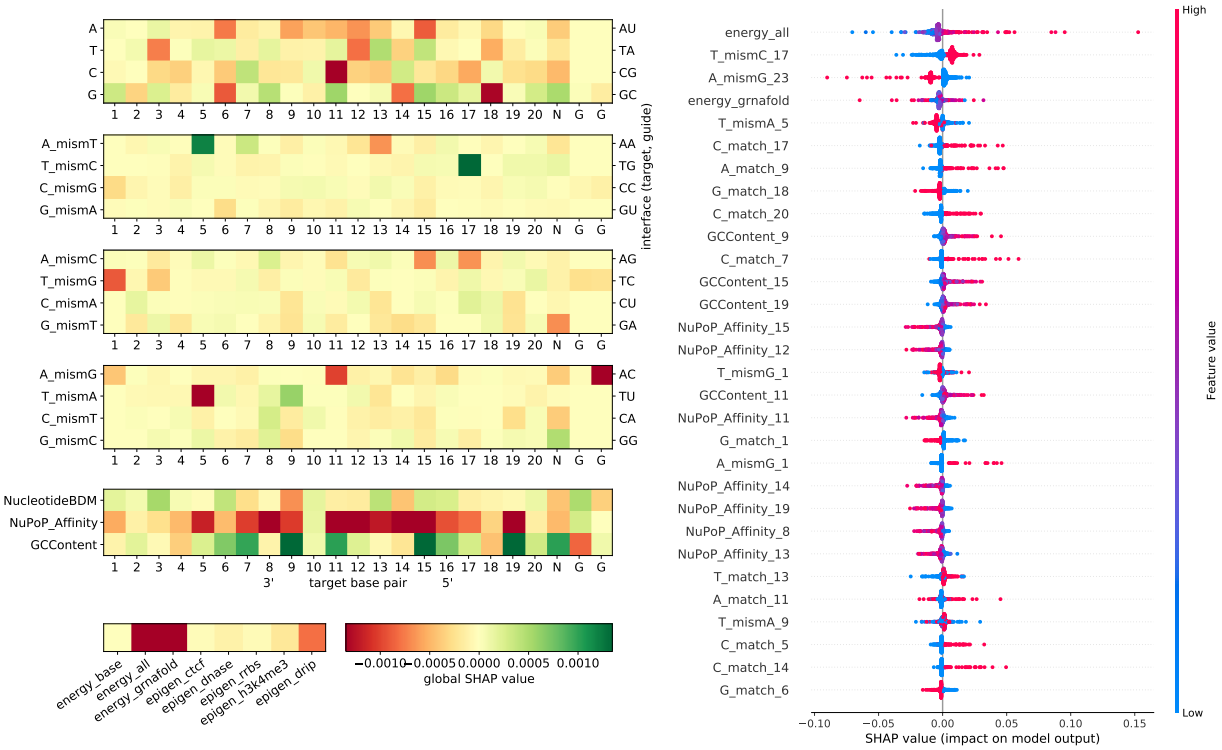


Figure 3.12: Base-pair resolved global SHAP values for the CNN_S5E2 classification model, where the global SHAP value of a given feature is defined as the mean SHAP value across 500 randomly sampled test set points. Test data consisted of the held out studies [90, 96, 97] from the crisprSQL dataset. Mismatch channels (middle three heatmaps) are represented by the (off-)target and on-target protospacer nucleotides (left vertical axis) as well as the physical base pair interfaces (right vertical axis), such that `A_mismT` describes all configurations in which an adenine on the target strand faces an adenine on the sgRNA. The bottom heatmap visualises the influence of a set of nucleosomal organisation features on cleavage activity. Negative global SHAP values (red) indicate an average predicted decrease in guide activity for the respective feature. A bar representation of this can be found in Figure B.2.

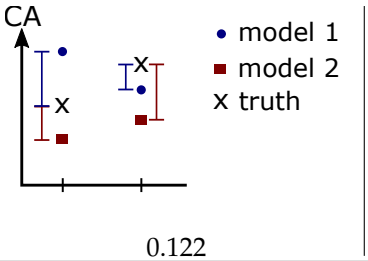
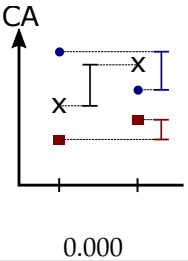
		$(\text{pred}_i - \text{truth}_i)^2$ prediction MSE	$\left(\frac{\text{pred}_i - \text{pred}_j}{\text{truth}_i - \text{truth}_j}\right)^{\frac{1}{3}}$ rel. pairwise difference
CRISPR-Net	sequence-only		
		0.122	0.000
piCRISPR	sequence-only	0.139	0.000
	sequence + epigenetic	0.085	0.140
	sequence + nucleosomal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

Table 3.2: Quantities gained on the subset of duplicate guide-target sequence pairs (testing scenario 3) using our piCRISPR CNN_S5E2 regression model and the CRISPR-Net model available online [67]. Schematic depictions for the calculation of benchmarks (MSE, relative pairwise difference) are shown. Predictions of two different models, as well as ground truths for two different data points are indicated by ■, • and ×, respectively. For piCRISPR, we have weakened the model by setting either the literature-standard epigenetic features (DNase, H3K4me3, CTCF, RRBS) or our newly utilised, context-based nucleosome positioning-related features (NuPoP Affinity, Nucleotide BDM, GC count) to a default value across for data points. The non-environmentally sensitive energy and sequence features were retained throughout. **Left column**: mean squared error (MSE) between predicted cleavage score and ground truth cleavage activity (CA), averaged over all groups of identical guide-target sequence pairs. **Right column**: How faithful a model is to the differences in biological environment for a given pair within such a group is measured by the average proportion of the true cleavage activity difference which the model predicts. This is zero for purely sequence-based models and unity for an ideal predictor. In order to emphasise small deviations which preserve the rank of predicted cleavage activities, we use the cubic root as a sign-preserving nonlinearity and term this quantity *relative difference*. The distributions of MSEs and relative differences are shown in Figures 3.13 and 3.14.

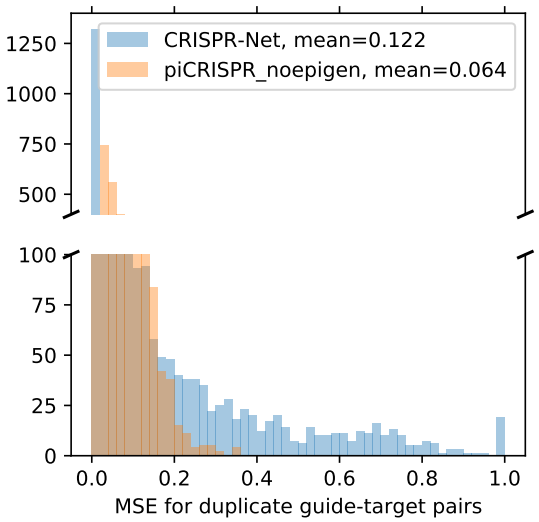


Figure 3.13: Distributions of mean squared error between cleavage activity and prediction score for the CRISPR-Net algorithm (blue bars) and our CNN_S5E2 model (orange bars), where in the latter case the literature-standard epigenetic features (DNase, H3K4me3, CTCF, RRBS) have been set to a default value across data points. The distributions for other sets of features can be found in Figure 3.14.

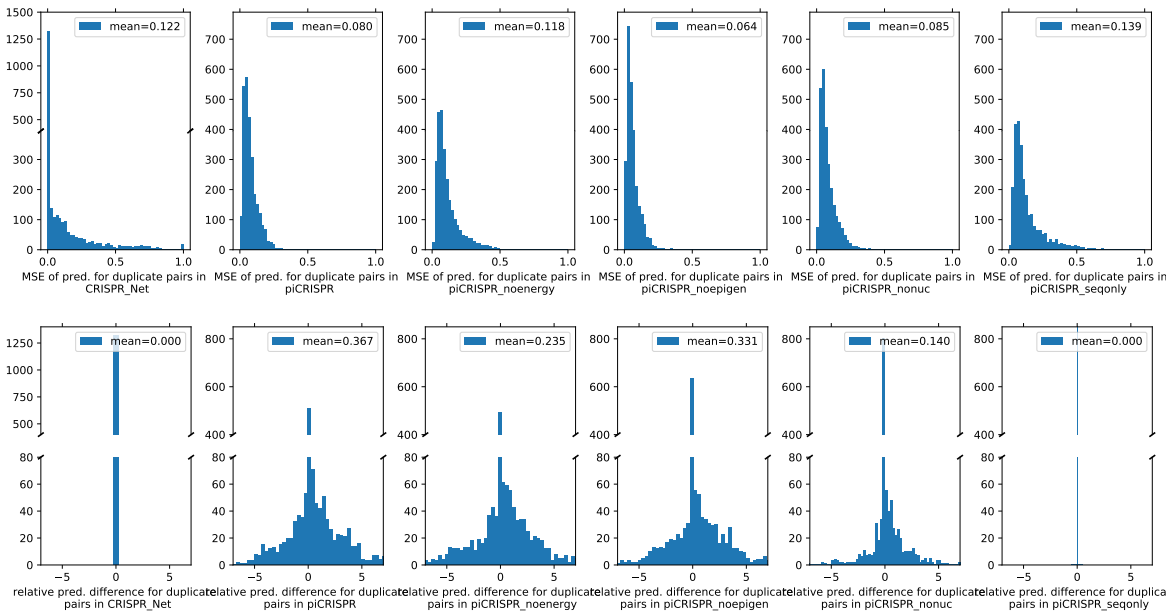


Figure 3.14: Distributions underlying Table 3.2.

3.3 Optimised selection of architectures, feature sets and encodings

Starting from the learnings above, we now employ two different feature encoding schemes which occupy different points in the tradeoff between sparsity and interpretability. The first was introduced in [67] and consists of one-hot encoded representations of the guide and target sequence which have been combined using a bitwise OR operation. In order to make up for the loss of information that this operation causes in terms of mismatches, two additional channels are added containing information about the directionality of the bases involved in the mismatch, i.e. which of the two entries describes the target and which the guide nucleotide. Note that the guide nucleotide is first translated into its corresponding target protospacer nucleotide. We call this encoding the 6×23 encoding based on the resulting shape of the sequence matrix (see Figure 3.15A).

Based on the energy-driven nature of binding and cleavage, we hypothesise that mismatched interfaces affect binding in a totally different way than matched interfaces. This has so far not been recognised in detail by off-target prediction algorithms. Since the 6×23 encoding contains the information about the precise nature of a given interface only implicitly, we decided to include a further encoding which does so explicitly. This uses a one-hot representation using the 16 letter cross product between guide and (off-)target nucleotide, and is hence termed the 16×23 encoding (see Figure 3.15B). This is similar to the encoding scheme in [65].

Both matrices are then concatenated with a matrix of base-pair resolved nucleosomal features and the `CRISPRoff` value [13] of the given target-guide interface repeated along the sequence axis. Exploring latent representations of guide or target is not within the scope of this work, given that it further complicates comparison between models.

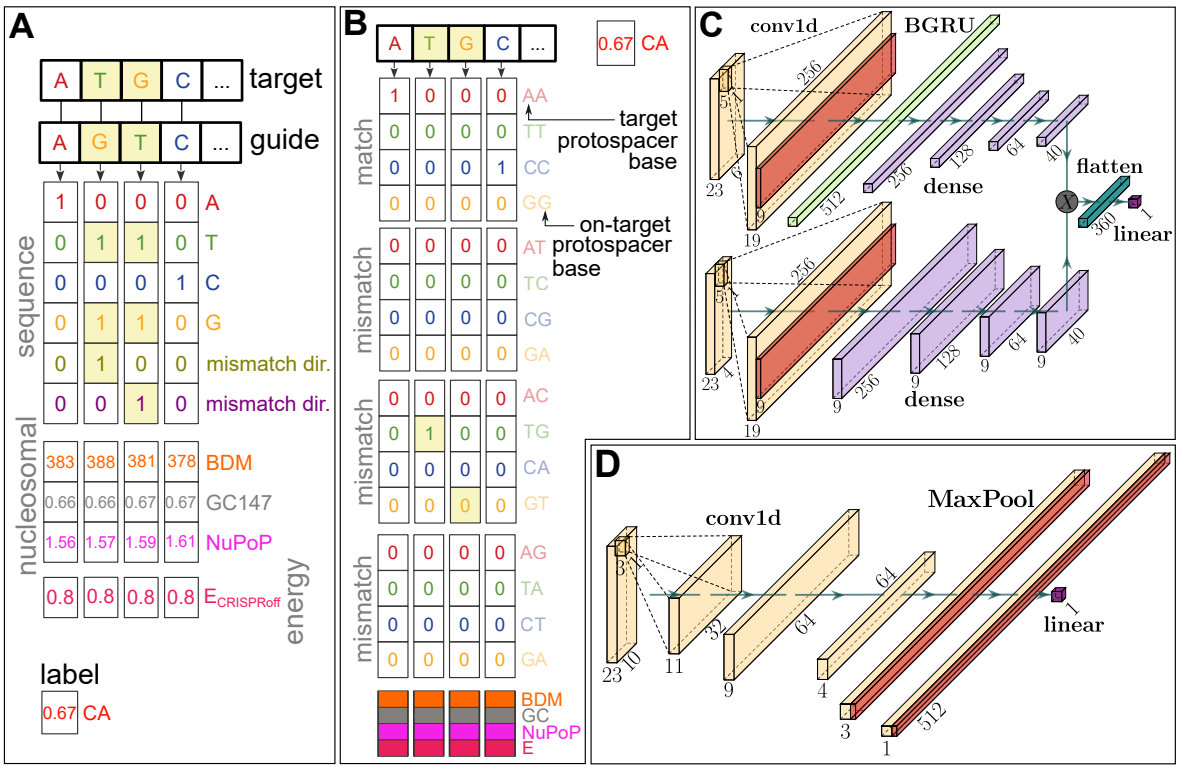


Figure 3.15: Overview of our two chosen feature encodings and models. **A** 6×23 encoding: (Off-)target and guide (on-target protospacer) sequences are one-hot encoded and copied together using a bitwise OR operation. In order to make this encoding lossless, two channels are added that encode which nucleotide is on the guide and which on the target, respectively (in the case of a mismatched interface). This is identical to the encoding used in [67]. The nucleosome positioning channels (147 bp GC count, Nucleotide BDM [131], NuPoP Affinity [128]) of the target as well as the `CRISPRoff` free energy estimate score [13, 118] are concatenated to this matrix. The Box–Cox transformed cleavage activity (CA) is used as a label. **B** The 16×23 encoding uses a 16-letter alphabet which explicitly contains information about the precise nature of the mismatch. For an explanation of panels **C** and **D**, we refer to Figure 3.2

3.3.1 Model Training & Evaluation

Model architectures, training hyperparameters and the retrieval of feature importance values remain as in the previous section. In order to allow for a yet more realistic literature comparison, we adjusted the testing scenarios to the following:

Testing Scenario 1: held out studies

In this scenario we hold out studies [90, 96, 97] from the training set. These studies have not been included in the training set for the state-of-the-art off-target prediction algorithm CRISPR-Net [67], such that they remain an independent test set to compare CRISPR-Net and piCRISPR side by side. The inherent class imbalance in this test set is 1:103.96. 22 % of the unique guides within the training set have at least one corresponding guide in the test set with five or fewer mismatches, indicating a satisfactory independence between training and test set.

Testing Scenario 2: literature comparison

In this scenario we use the CIRCLE-seq [30] dataset as the held out test set, as was done in [68]. The exact test set has been replicated using the code provided by the authors, such that comparison values could be taken straight from publication [68]. Nucleosomal and empirical energy data was filled in using the crisprSQL dataset.

Testing Scenario 3: set of duplicate pairs

This scenario remains as detailed in section 3.2.

3.3.2 Results & Discussion

Testing Scenario 1

Figure 3.16 shows the regression and classification performance of our piCRISPR-implemented models, with the 6×23 RNN model yielding the highest benchmarks. Adding the nucleosomal features (models labelled "nuc") considerably improves model performance according to all benchmarks, supporting our hypothesis that nucleosomal features can serve as a key ingredient to cleavage prediction. It should be noted that almost all models achieve an area under ROC curve of 0.95 or above, meaning area under PRC curve is a more suitable bench-

mark for state-of-the-art cleavage activity classification models. The predictive advantage of piCRISPR over literature algorithms is most pronounced for this benchmark.

A direct comparison with prediction results obtained from the published versions of CnnCrispr, CRISPR-Net and DeepCRISPR on the identical held out test set (bottom panel of Figure 3.16) shows that piCRISPR achieves higher classification benchmarks in terms of areas under ROC and PRC curve for all three individual studies contained in the test set, except for study [90] for which piCRISPR and CRISPR-Net achieve comparable benchmarks.

Testing Scenario 2

Figure 3.17 shows that when testing on the CIRCLE-seq dataset [30], piCRISPR performance drops slightly as compared to testing scenario 1. Especially RNN models generalise slightly worse to this study. We still observe that nucleosomal features enhance the performance of a model, and piCRISPR still outperforms both CRISPR-IP and CRISPR-Net models according to area under PRC curve.

Testing Scenario 3

As for the models from section 3.2, Table 3.3 shows that the performance of our improved set of models, measured by the mean squared error of predictions within a group of data points that share both guide and target sequence, is considerably improved by introducing features beyond sequence information (left column). The resulting distribution of MSEs is shown in Figure B.8.

Looking at the relative pairwise difference, we observe that introducing features beyond sequence leads to an increase of the average proportion of true cleavage frequency differences between points of differing biological environment which is captured by the model. This is true for both DeepCRISPR and piCRISPR. Whilst the full feature set in piCRISPR achieves the highest proportion in comparison, it is the piCRISPR sequence-only model that achieves the lowest overall mean squared error. This indicates that a low mean squared error

does not necessarily go hand in hand with the model drawing the correct conclusions from environmentally sensitive features. This can be seen as well when considering the comparably small relative pairwise difference that is recovered by the DeepCRISPR model from the literature-standard epigenetics channels to which it has access. Figure 3.18 visualises the distributions of the relative pairwise difference quantity for DeepCRISPR and our 6×23 model. We can observe that the distribution for our model is markedly skewed towards positive values, indicating that nucleosomal features on average serve to explain the differences in cleavage activity resulting purely from the biological environment of the cut site. This effect is considerably less pronounced for the DeepCRISPR prediction, indicating that literature-standard epigenetic features are on average less well suited for this task.

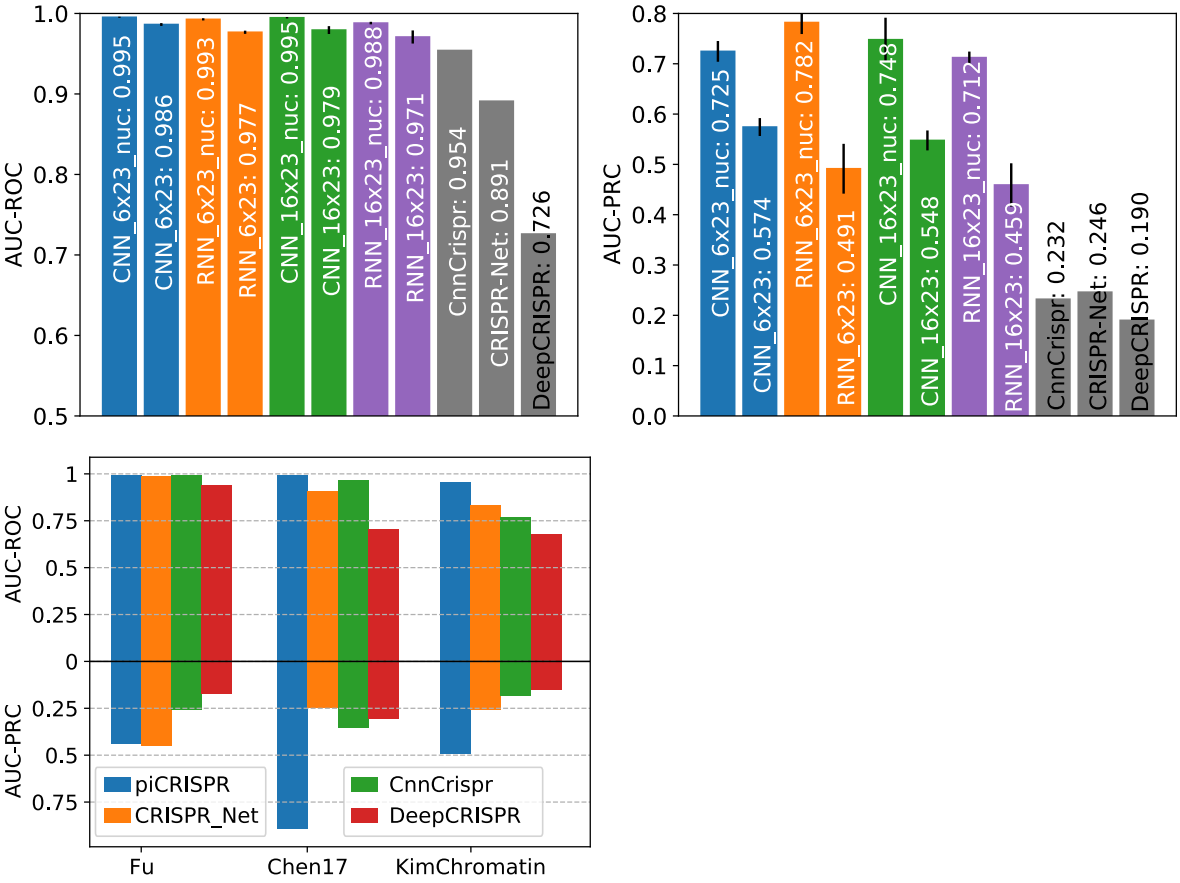


Figure 3.16: Comparison of piCRISPR models with published algorithms. All models were tested on held out studies [90, 96, 97] (testing scenario 1). Non-validated data points have been oversampled in the test set to match the class imbalance of 1:79.35 found in the dataset I-1 from [67]. piCRISPR models have been trained on the remaining data points within the crisprSQL data set. **Top two plots**: Comparison with three published off-target prediction algorithms [65, 67, 75] that were run on this test set. Within a model family of the same colour, the model labelled "nuc" contains nucleosomal features whereas the other does not. piCRISPR training and testing have been repeated 5 times to obtain mean and standard deviation as shown. For the underlying ROC and PRC curves see Figure 3.10. **Bottom plot**: AUC-ROC and AUC-PRC benchmarks for the RNN 6×23 model with nucleosomal features, resolved by individual study within the held out test set.

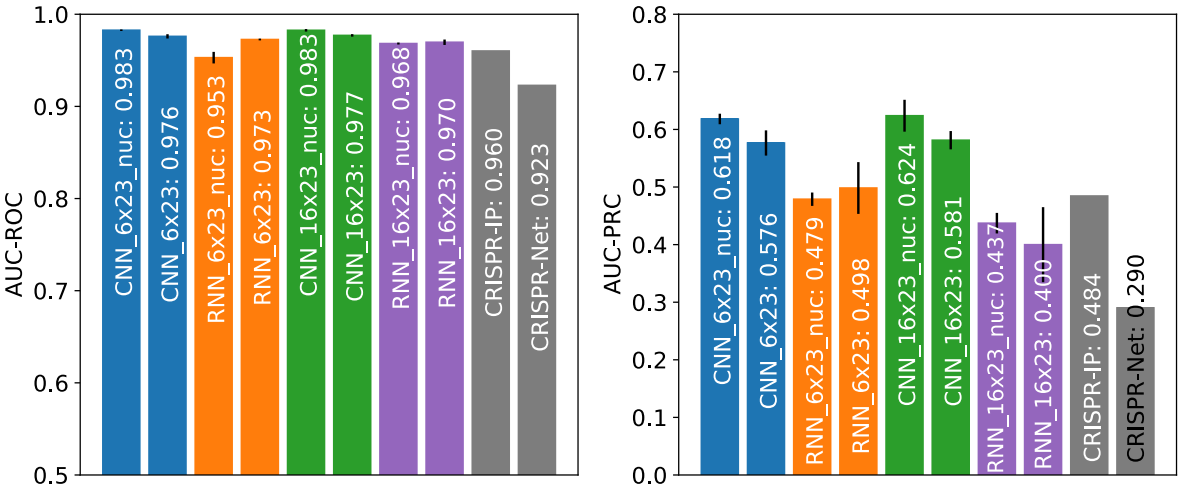


Figure 3.17: Comparison between piCRISPR, CRISPR-Net [67] and CRISPR-IP [68], with comparison values for the latter two models taken from [68]. All models were tested on the identical subset of the CIRCLE-seq dataset [30] which was used in [68] (testing scenario 2). Note the slightly reduced performance of the RNN models compared to testing scenario 1 (Figure 3.16).

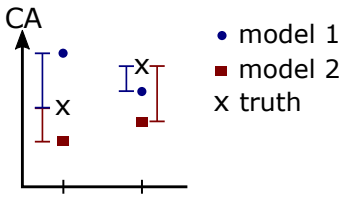
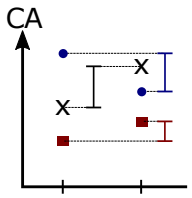
		$(\text{pred}_i - \text{truth}_i)^2$	$\left(\frac{\text{pred}_i - \text{pred}_j}{\text{truth}_i - \text{truth}_j}\right)^{\frac{1}{3}}$
		prediction MSE	rel. pairwise difference
CRISPR-Net	sequence-only	 0.157	 0.000
DeepCRISPR epigenetic		0.019	0.018
piCRISPR	sequence-only	**0.012**	0.000
	full feature set	0.101	**0.349**

Table 3.3: Benchmark quantities gained on the subset of duplicate guide-target sequence pairs (testing scenario 3) using our 6×23 CNN model as well as the CRISPR-Net [67] and DeepCRISPR [75] models for comparison. This is an updated version of Table 3.2. For a precise description, see there. All underlying distributions are shown in Figure B.8.

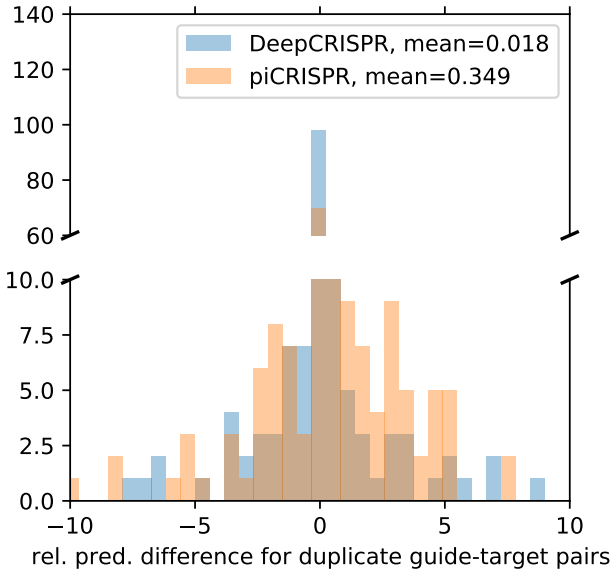


Figure 3.18: Distributions of the relative pairwise difference as defined in Table 3.3 for the DeepCRISPR algorithm (blue bars) and our 6×23 CNN model (orange bars). The distributions for other sets of features can be found in Figure B.8.

Feature importance

Due to its comparatively stronger prediction benchmarks between testing scenarios 1 and 2, we use the 16×23 CNN classification model in testing scenario 1 to extract feature importance values of unprecedented resolution.

Figure 3.19 shows that, similar to what we observed in the previous section, this model draws on sequence features which stem from mismatched interfaces differently than on those from matched interfaces. As was the case there, global SHAP values suggest that the preference of the variable PAM nucleotide at position 21 is contingent on the specific sgRNA–DNA interface formed. We recover the preference for cytosine at position 17 [75, 133, 136] as well as position 20 [75, 108, 133] found in literature for matched interfaces. However, for mismatched interfaces, cytosine is disfavoured. Whilst we cannot recover a strong preference for the variable PAM nucleotide at position 21 for matched interfaces, we observe the preference for guanine reported in literature [75, 108, 136] for mismatched interfaces. As was the case in Figure 3.12, this supports the notion that a concentration on guide-target interfaces rather than pure base identities is necessary for off-target prediction. Again, deeper insight is required than the notion of a preferred base at a specific position, necessitating the consideration mismatch interfaces together with sequences in the desired genome, not just the sequence of the putative guide, for sgRNA design.

The model draws heavily on the empirical energy estimate feature $E_{\text{CRISPRoff}}$ which yields the largest global SHAP value. We further observe a considerable correlation between its value and the SHAP value attributed to it by the model (Figure 3.21). An energy score of $E_{\text{CRISPRoff}} = -1.15$ has a neutral influence on cleavage activity in the 16×23 CNN model, with higher (lower) values yielding larger (smaller) SHAP values, i.e. a positive (negative) influence on cleavage activity. Small free energies of the DNA-RNA heteroduplex appear to favour cleavage activity, whilst very strong binding is not optimal for cleavage activity.

When considering nucleosome positioning-related feature channels, we see that the 147

bp GC content around each nucleotide has a net positive influence on cleavage activity. Similar to the argument in [124], this can be attributed to the increased bendability of GC-rich DNA [141] which is beneficial to Cas9 sequence readout during binding [142]. We further observe that the NuPoP Affinity score ranks higher in terms of global SHAP value than most sequence features. The negative influence of nucleosome affinity can be explained by the reduced accessibility of high-affinity DNA regions, and is observed strongly between nucleotides 5 and 19. This effect has been observed in [124] as well. We further observe an overall negative influence of low Nucleotide BDM values on cleavage activity, supporting what has been observed in preliminary, non-sequence based models in [124]. For this model, the effect of Nucleotide BDM is markedly smaller than for NuPoP Affinity. We attribute this to the fact that Nucleotide BDM has been shown to strongly correlate with nucleosome occupancy [131], i.e. a time-averaged probability for a sequence to be bound within a nucleosome, whereas NuPop Affinity yields a probability for nucleosome positioning on longer timescales. We hypothesise that in the former case, CRISPR machinery can have access to the site during times when the sequence is exposed, making nucleosome occupancy a less good predictor for accessibility than nucleosome positioning. The NuPoP Affinity and GC content features further show a distinct 2 bp periodicity, with importance magnitude and in the case of GC content even sign alternating every other base pair, especially between nucleotides 1 and 17. We conjecture that this is due to our chosen uneven convolutional kernel size 5.

A similar analysis for the 6×23 CNN model can be found in Figure B.5 and for the 16×23 RNN model in Figure B.7. Note that within the nucleosomal feature class, the RNN models attribute a higher importance to the Nucleotide BDM feature than the CNN models scrutinised here. This could in part explain their slight difference in performance between testing scenarios 1 and 2.

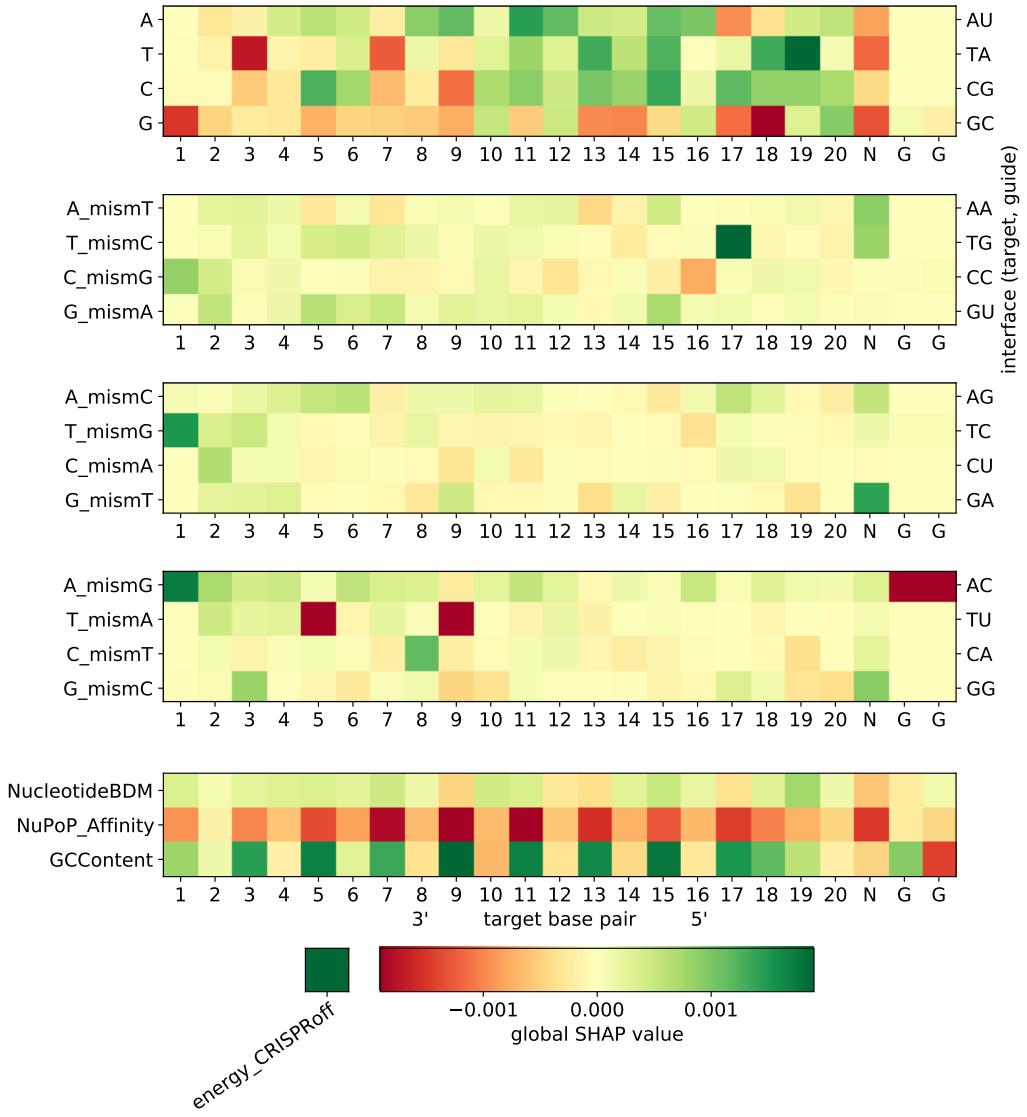


Figure 3.19: Global SHAP values for the 16×23 CNN classification model. Negative global SHAP values (red) indicate an average predicted decrease in guide activity for the respective feature. Mismatch channels (middle three heatmaps) can be represented by the (off-)target and on-target protospacer nucleotides (left vertical axis) as well as the physical base pair interfaces (right vertical axis), such that `A_mismT` describes all configurations in which an adenine on the target strand faces an adenine on the sgRNA. The bottom heatmap visualises the influence of our chosen set of nucleosomal organisation features on cleavage activity. A bar representation of this can be found in Figure B.6.

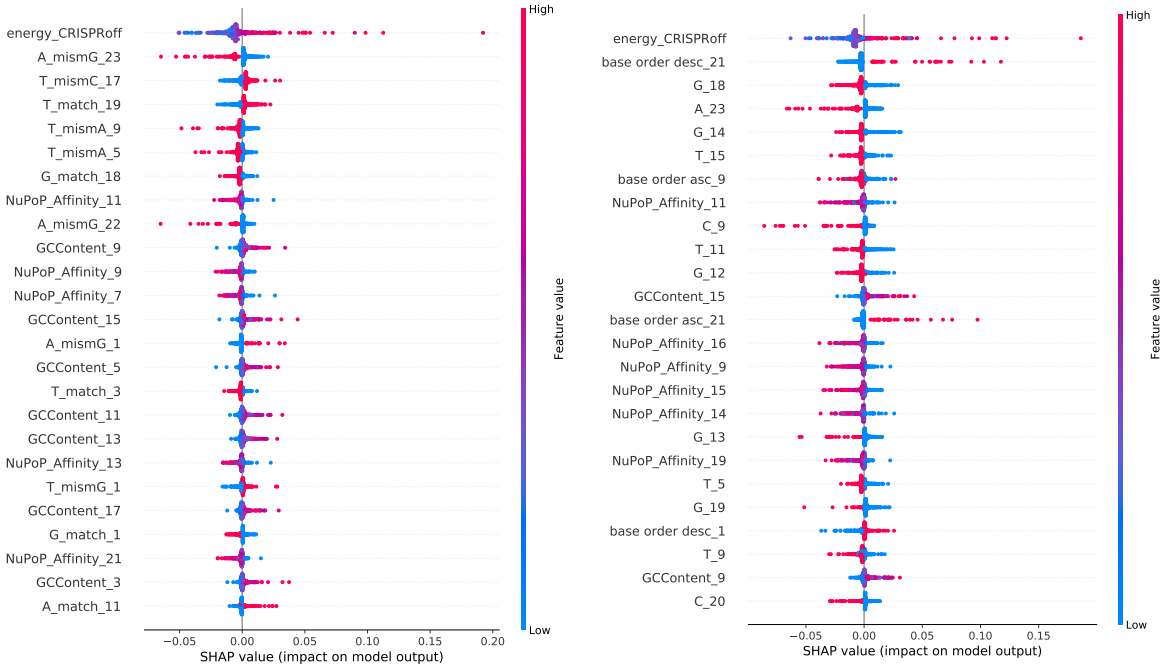


Figure 3.20: Base-pair resolved global SHAP values for the 16×23 (left panel, see Figure B.6) and 6×23 (right panel, see Figure B.5) CNN classification models. SHAP values have been obtained on the held out studies [90, 96, 97] from the crisprSQL dataset. Note that high values of the NuPoP Affinity feature (red dots), i.e. highly positioned nucleosomes, always influence the model towards reduced cleavage activity (negative SHAP value).

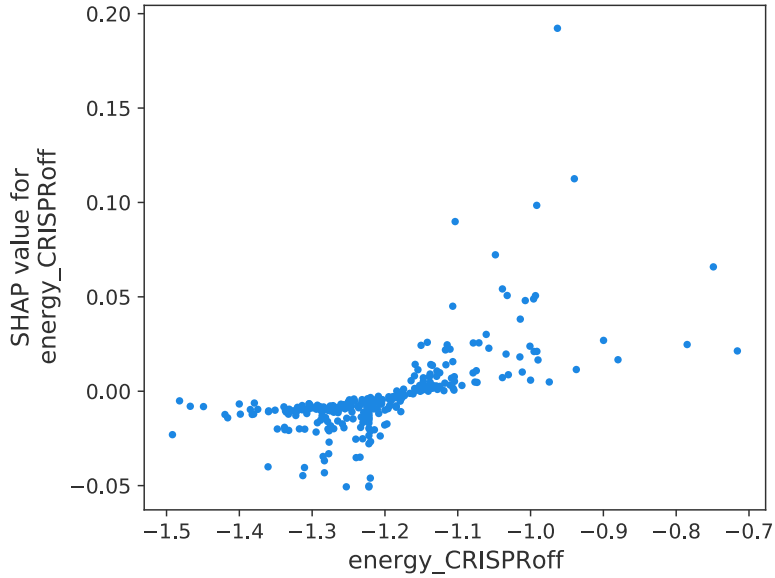


Figure 3.21: SHAP dependence plot for the $E_{\text{CRISPRoff}}$ feature. The horizontal axis is in arbitrary units as provided by CRISPRoff [13]. Data has been gained using the 16×23 CNN classification model (Figures 3.20 and B.6). For energy scores $E_{\text{CRISPRoff}} > -1.15$, the RNA-DNA heteroduplex is weakly bound and opens easily, hence favouring cleavage (positive SHAP value), whereas for smaller $E_{\text{CRISPRoff}}$ it becomes more tightly bound.

3.3.3 Optimal split of data into train & test set

In order to achieve a reliable assessment of model generalisation performance, Xiang *et al.* recommend the splitting of guides into train and test set based on minimising the pairwise Hamming distance of guides between these two sets [143]. They consider two guides with a Hamming distance of ≤ 4 *close* and split close pairs into different subsets. Subsequently, they grow these subsets by recursively adding further close guides to the guides in each set. Guides not close to any other guide were assigned to subsets pseudorandomly, aiming at achieving a balance of high, medium and low activity guides between the subsets.

Due to the iterative nature of this process, we conjecture that this might not lead to a close to optimum split of guides into separate subsets with the aim of minimising the pairwise Hamming distance between guides from each pair. Furthermore, we argue that, applying the idea from Xiang *et al.* strictly, parameters such as cell line, the type of DSB quantification assay and the delivery method of the CRISPR machinery into the cell ought to be taken into account as well as the guide sequence. Ensuring a maximum distance of all these parameters between train and test set then truly tests a model's generalisation performance across their different combinations. We therefore suggest two alternative approaches to come closer to this aim, and evaluate the generalisation performance of a model trained with the respective split. On this data, we want to base a discussion of the relevance and necessity of such an engineered data split.

Max cut approach

Data about cell line, type of DSB quantification assay and CRISPR delivery method was taken from the crisprSQL dataset as described in section 2.2.1. We employ a categorical encoding of these three parameters and a one-hot encoding of the guide sequence of all human-origin data points within crisprSQL and perform a density-based DBSCAN clustering [144] on this data set. With an empirically optimised set of parameters, we obtain 83 clusters of points and

calculate the average pairwise Hamming distance between points in either set (see Figure 3.22). This triangular matrix is fed into an implementation of the max cut algorithm, which searches for an allocation of the dataset into two subsets that maximises the average pairwise Hamming distance of the pairs of two guides, one from each subset. We use the Python library `maxcut` for this. This process is repeated with different random seeds until we find an allocation that results in a split that is close to an 80/20 ratio in terms of number of data points in each subset, as well as offers a roughly similar ratio between measured and augmented data points in both subsets.

Greedy approach

In order to guarantee that no guide appears in train as well as test set, we allocate all data points of a given guide into either train and test set, again ensuring that the resulting overall split is close to an 80/20 ratio and has roughly similar ratio between measured and augmented data points in both subsets. This approach only takes only the guide sequence into account.

For both the greedy and the max cut approach, we additionally run training and testing with a small admixture between train and test set in order to observe the effect of the approaches in transition into a random train/test set allocation. The results of this are shown in Figure 3.23.

Results & Discussion

Figure 3.23 shows that the max cut split results in a drastically reduced model performance, with the model practically deteriorating to a random classifier. We conjecture that this is due to a clean split between cleavage assays, delivery methods, cell lines and guide sequences into train and test set, which reduces the train set diversity below a level that would represent the world based on which the model will be asked to make predictions. It is in this sense an unnatural and unrealistic requirement for the model to generalise under these circumstances,

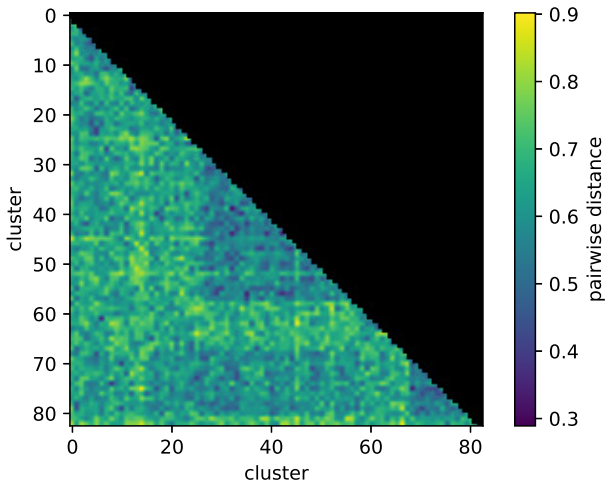


Figure 3.22: Visualisation of the Hamming distance matrix between the clusters identified by the DBSCAN clustering [144] of guides based on their sequence, cell line, DSB quantification assay and delivery method. This matrix is fed into a max cut solver in order to find an allocation of data points into train and test set that maximises the average Hamming distance between pairs of data points, one from each subset.

questioning the intuition by Xiang *et al.* [143]. It is worth noting that the average number of close test guides per train guide in this approach is still close to one in the max cut approach, indicating that including non-sequence based features in the max cut consideration led to a somewhat imperfect separation between close guides.

This point is remedied in the greedy split approach, which by construction splits apart close guide sequences cleanly into train and test set. For this split, we find that classification benchmarks suffer less drastically than for the max cut approach, with the greedy split still achieving an area under PRC of 0.4. This again supports the notion that the diversity of non-sequence parameters is more crucial for model generalisation than the proximity of guide sequences between train and test set.

We find that when inverting the allocation for only 2% of data points, a process we term *admixture*, we recover the generalisation performance of the model with random 80/20 allocation of data points into train and test set, respectively. We see this as an indicator that

there is an observable effect of the similarity of guide sequences between train and test set leading to an overestimated generalisation performance of the classifier. However, this effect vanishes with even a small number of close data points between train and test set through random admixture. A small-proportion overlap like this would be expected in a natural use of the trained model as well. Furthermore, the effect is reduced for our literature comparison using testing scenario 2 in which a set of complete studies with a distinct set of guides was held out as a test set. We therefore argue that a careful separation of guides into train and test set can aid but is not crucial in obtaining meaningful modelling benchmarks.

Following the induced DNA double strand breaks, cell-internal machinery as described in section 1.1.3 is activated for repair. Cleavage is a necessary consideration when assessing the security of CRISPR/Cas-induced editing outcomes, but its subsequent repair determines the genomic impact. We thus now turn our attention to understanding how certain DSB repair mechanisms determine these outcomes with the aim of predicting outcome distributions.

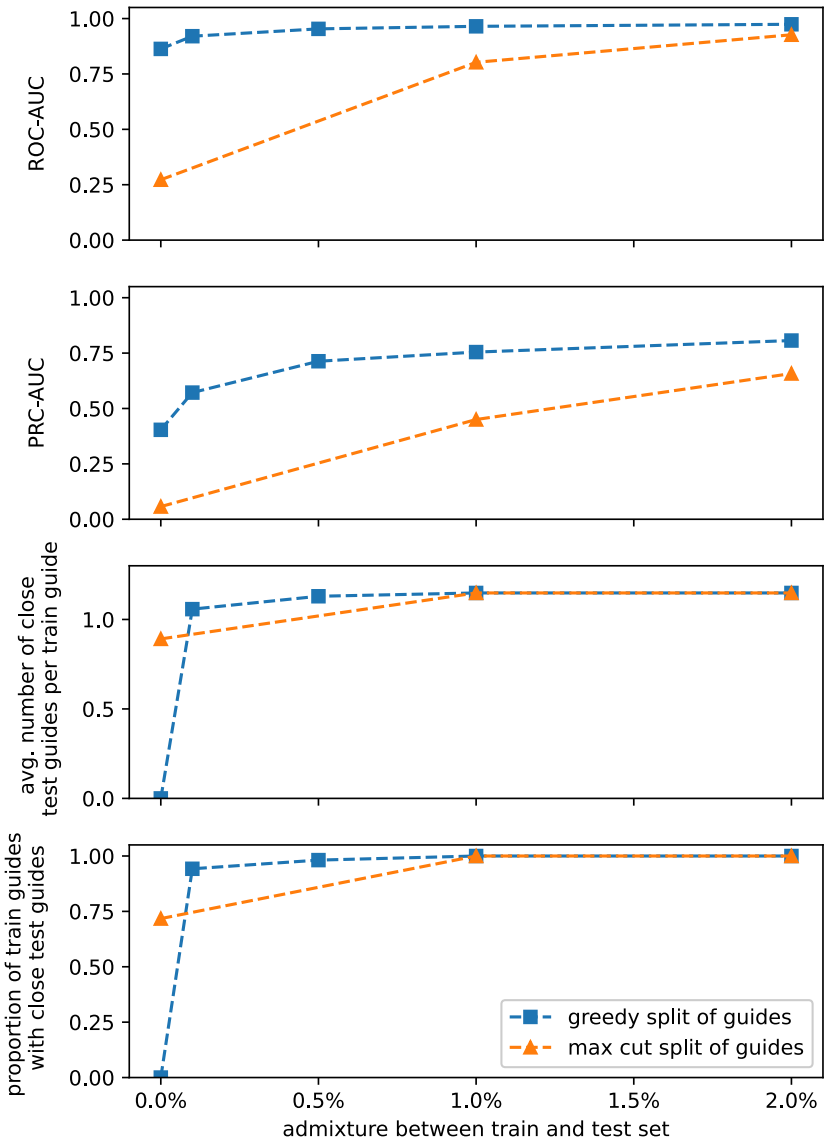


Figure 3.23: Performance benchmarks of the 16×23 RNN classification model in testing scenario 1 (top two plots) and quantities characterising a train/test split (bottom two panels). The horizontal axis refers to a random admixture between the two sets, i.e. moving towards a random allocation of guides to train and test set. We show this data for both the max cut and the greedy approach described above. Dashed lines are given as a guide to the eye.

Chapter 4

CRISPR repair outcome prediction using structural calculations

After a CRISPR-induced double strand break, cell-internal repair machinery is recruited to mend the open-ended DNA. This can create a variety of outcomes, often resulting in mutagenesis through insertions or deletions around the cut site. An assessment of CRISPR-based gene editing safety must therefore consider this step as well.

In this chapter, we present our approach to increasing the understanding of the outcome of endogenous DNA repair pathways through the use of structural thermodynamic calculations.

4.1 Introduction

Since its discovery in 2012 [23], CRISPR-Cas9 technology has transformed the field of DNA editing [4, 16, 145]. A 23bp long synthetic guide RNA is used to guide the Cas9 enzyme to a desired locus on the genome, where it subsequently induces a double-strand break (DSB). Cell-internal repair machinery is recruited to DSBs, resulting either in perfect repair of the site, or mutations which might induce frameshifts, thereby preventing the correct translation of the respective protein.

It has been shown that the variation in editing outcome occurs in a non-random way and is therefore reproducible [76, 146, 147]. Several datasets have been generated today that give the distribution of DSBs and resulting editing outcomes [76, 77, 148]. Starting from these, efforts to model outcome distributions have resulted in six prediction models, an overview of which can be found in [80]. These all use varying forms of sequence encoding as input features.

In this study, we want to contribute a fundamentally new approach to modelling outcome distributions by injecting physical knowledge about the underlying repair processes in the form of computationally estimated free energies. Due to its large size and accessibility, we utilise the FORECasT dataset [76] for this.

4.2 Preliminary considerations

In order to estimate the viability of a large-scale outcome prediction model based on calculated thermodynamic quantities, we scrutinise whether we can rationalise post-cleavage outcomes with the computational tool MOSAICS [49] which has been developed in our group.

To this end, we have devised two approaches, the first of which yields an energy difference in dependence of indel length, whereas the second yields an energy difference in dependence of indel position across a sequence defined in the experimental dataset (often called an oligonucleotide or short *oligo*). For both approaches, we locally construct right-handed B-DNA pdb files with the respective sequence and base pair per turn ratio using the `fd_helix` package from Nucleic Acid Builder [149]. These can then be fed into MOSAICS sequentially, which runs an atomistic conjugate gradient based energy minimisation on the molecule using the amber99-bs0 empirical force field [45, 150]. The DDOS distance-dependent dielectric [151] was used in order to implicitly represent the solvent without having to add water molecules explicitly, using the default parameter values [150]. This approach has been shown to agree well with experimental data.

Once the energy gradient falls below a preconfigured threshold[1], MOSAICS yields the total potential energy of the molecule in kcal/mol. For a sample MOSAICS output see Appendix C. In order to allow for data generation in an acceptable time frame, calculations were conducted in parallel using 10 workers on an Intel i5-6500 3.2 GHz CPU.

4.2.1 Insights from the FORECasT dataset

The FORECasT dataset [76] has been generated by synthesising more than 40,000 gRNAs and their respective on-targets, introducing them to Cas9-expressing human K562-Cas9 cells using lentiviral transfection, and subsequent high-throughput sequencing of the target locus and its repair outcomes using a shared primer. The dataset comes with

- an index list of guides with their target context and cut site (we refer to this as an *oligo*),
- the measured outcome profile, one for each of these oligos, usually between 10 and 100 different outcomes with their normalised read count, parametrised by an *indel signature*, and
- a list containing the allocation of a given oligo to its sequence-identical counterpart in the other biological replicate, where the replicates differ in their gRNA scaffold.

An indel signature refers to a tuple consisting of indel length and indel starting position relative to the position of the CRISPR-induced double strand break (cut site).

For the purpose of this preliminary study, we used the oligo data in the folder *7312067/ST_April_2017_K562_800x_6OA_DPI3_Old7*. After parsing the length and absolute starting locus of each of these outcomes, omitting the remaining low-read tails once 80% of the total reads for each oligo had been included, we scrutinised the correlations between several parameters in the data set. The most insightful of these is depicted in Figure 4.1, where we begin to see a correlation between observed outcome frequency and indel length.

[1]The full MOSAICS configuration file was as in section 4.3 and can be found in the online repository mentioned there.

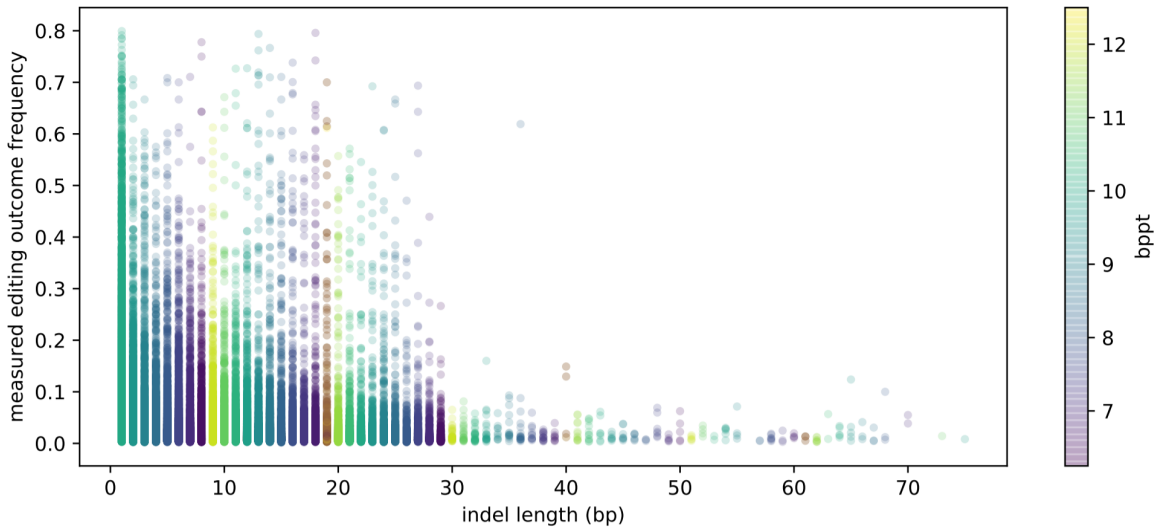


Figure 4.1: The measured editing outcome frequency for the first 50,000 outcomes across several thousand oligos in dependence of their indel length. The colour scale shows our proposed base pair per turn value (bppt) for each measured indel length (see below). Given that points are plotted with 20% opacity, we can begin to see a correlation between outcome frequency and indel length.

4.2.2 Approach 1: Adjust residue twist angle according to indel length

The hypothesis here is that an indel introduces torsional strain into the DNA (see insets in Figure 4.3), which then deviates from its standard value of 10.5 base pairs per turn [41] until it is unwound by topoisomerases. It therefore incurs an energy penalty $\Delta E_{\text{bppt}} = E_{\text{strained}} - E_{10.5}$. In order to obtain this, we calculate the energies of the editing outcome sequence both at a newly calculated residue twist angle (labelled by the corresponding number of base pairs per helix turn, bppt) and at the 10.5 bppt no-torsion configuration and plot the difference (figures 4.2 and 4.3). The bppt value is calculated under the assumption that the the time constant of the DNA torsion unwinding is such that during repair, the repair-induced torsional strain is distributed equally among the residues within a 21bp window around the cut site. We therefore assume a constant bppt value within this window and use it as a reference frame for our energy calculations. Furthermore, adding integer multiples of 10.5 bppt as well as halving the bppt within the considered window still results in a configuration for which no torsion remains outside the window. Using these two assumptions

limits the bppt values to the range of [6.25, 12.5].

It is worth noting that this approach assumes non-homologous end joining (NHEJ) as the DSB repair mechanism, even though the data reportedly contains a mixture of NHEJ and microhomology-mediated repair (MMEJ) outcomes. For the future, we plan to use data sources which differentiate between the two repair mechanisms.

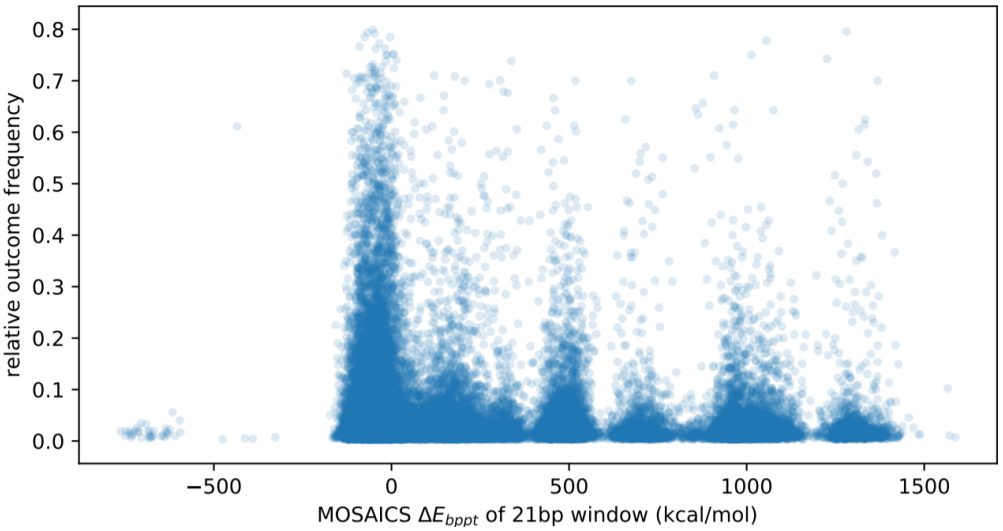


Figure 4.2: Measured indel rate plotted against the calculated potential energy difference between unstrained DNA and DNA which has been torsionally strained according to indel length for a 21bp window around the cut site (50,000 individual outcomes). High-frequency outcomes appear to cluster around favourable energy values (Spearman $r = 0.15$ for this unprocessed set), making this value suitable for informing a repair outcome predictor.

Plotting the above energy difference ΔE_{bppt} separated by base pair per turn value, we observe two distinct contributions (Figure 4.3): a *sequence-dependent* contribution which determines the width of the energy distributions of each individual bppt value, and a *torsion-dependent* contribution which shifts the center of energy distributions for each bppt value. The region between 8.5 and 10.25 bppt where these two contributions are comparable in size seems to be a good candidate feature to inform an *indel length* prediction algorithm, given that current models highlight the influence of sequence context on editing outcome. More and more editing outcomes would fall into this region when increasing the aforementioned

size of the base pair window for the energy calculation, although this goes at the expense of computation time.

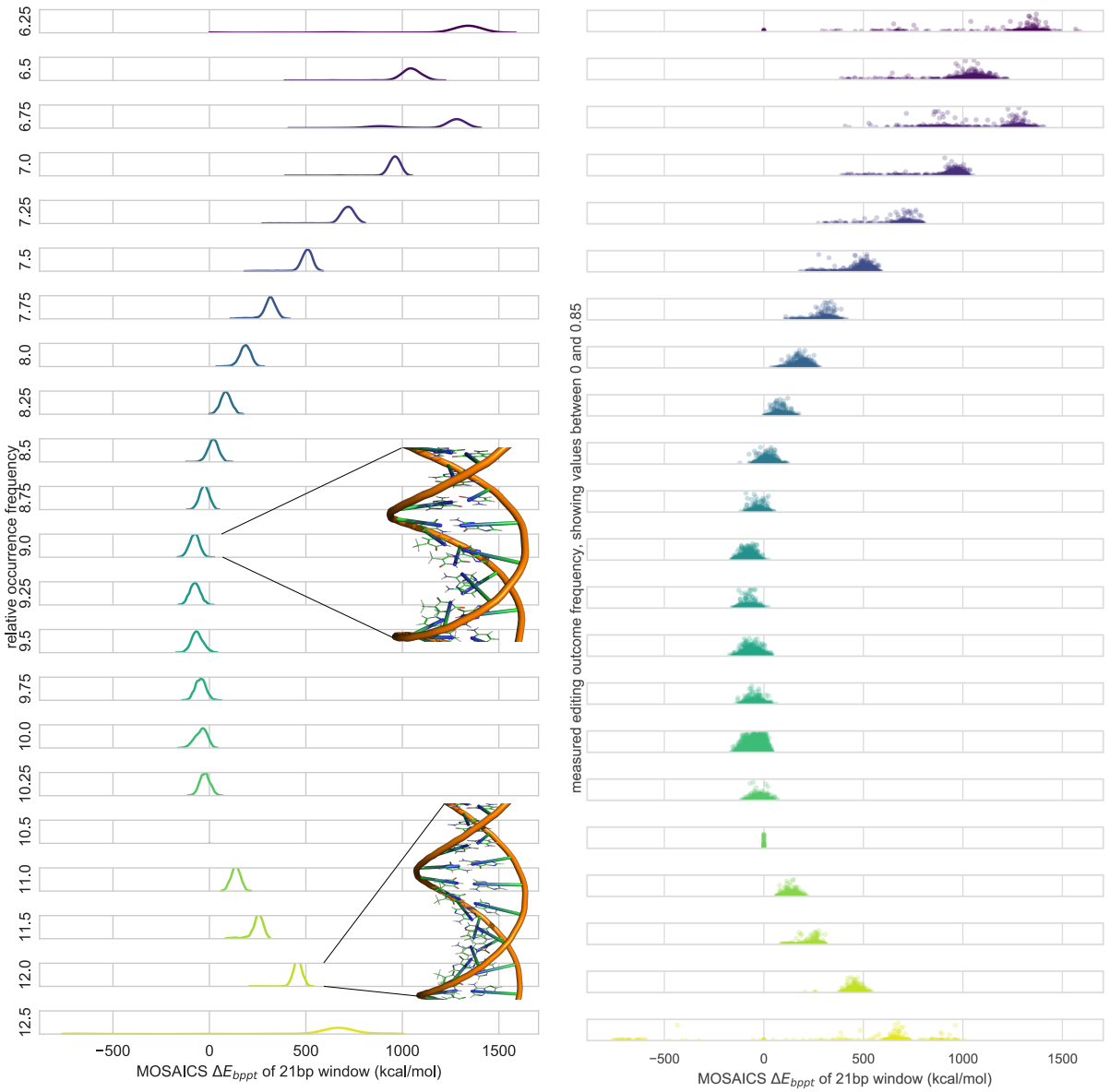


Figure 4.3: Left: Distribution of all the calculated ΔE_{bppt}, grouped according to the residue twist angle (bppt) templates used to calculate the respective values. Insets are DNA structures for a 9 and 12 base pairs per turn residue twise angle, respectively. **Right:** Measured outcome frequency distributions over the calculated potential energy difference. Note that the 10.5 bppt class mainly includes zero-length indels. Colours as in Figure 4.1.

4.2.3 Approach 2: Sweeping indel position across the cut site

Here, we calculate the energetic cost for the loss of base-stacking interaction energies associated with a given hypothetical repair outcome. This involves calculating the potential energies of four sequences (A, B, XZ, Y) contributing to this, and taking the difference $\Delta E = E_{xz} + E_y - (E_a + E_b)$ (see Figure 4.4).

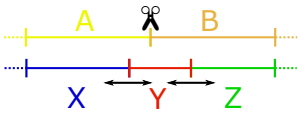


Figure 4.4: Sequences involved in calculating the loss of base-stacking interactions during the repair of a cut at the indicated site, resulting in the deletion of sequence Y. In Figure 4.5, we sweep the position of sequence Y relative to its measured value (black arrows) without changing the length of Y.

As Figures 4.5 and 4.6 suggest, this energy value could be a valuable feature for predicting the *position* of a fixed-length indel relative to the cut site, e.g. in a multi-task learning model where indel length and location relative to the cut site are predicted in separate stages.

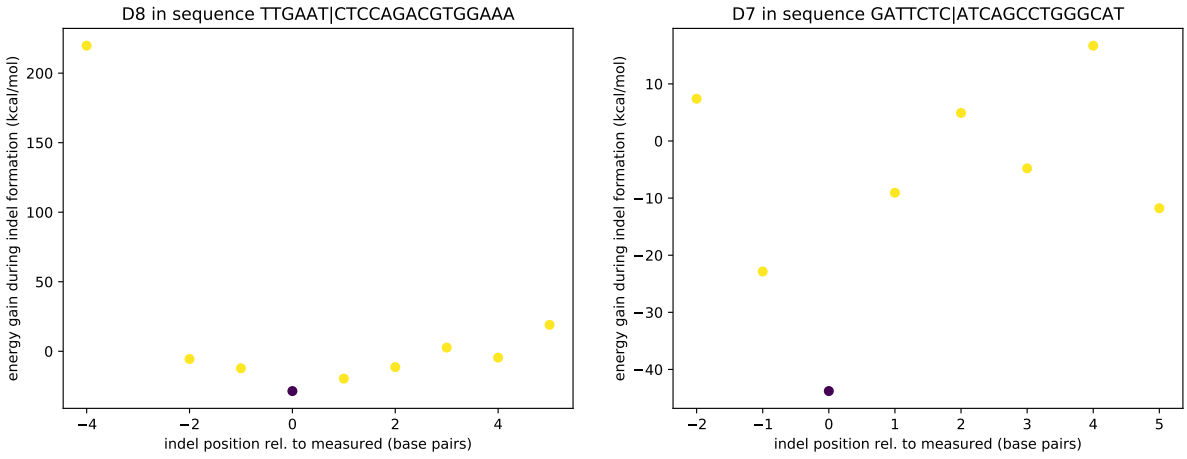


Figure 4.5: Resulting internal energy differences ΔE during outcome repair, sweeping the tentative indel position around the measured outcome (dark point) for two example indel data points. The cut location is always included in the sweep in order to mimic the process of NHEJ repair, which limits the sweep range. Figure 4.6 confirms that these are representative examples in the sense that for a fixed indel length, the measured indel position mostly sits at the bottom of all hypothetical indel positions.

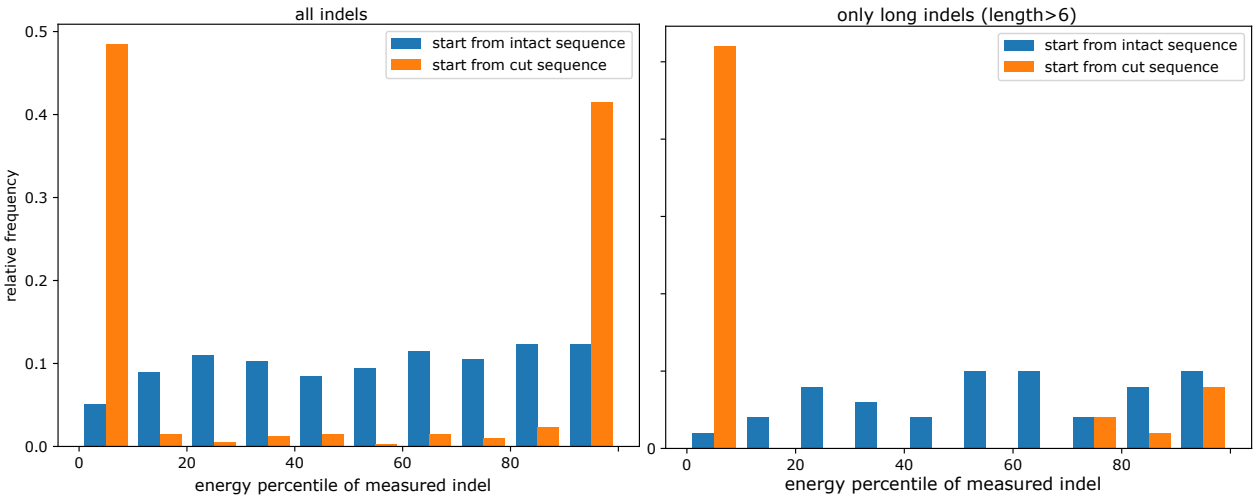


Figure 4.6: Histograms for the ΔE percentile in which the measured indel location falls. The plot on the right only considers long indels (length > 6), whereas the plot on the left considers all indels. For comparison, blue bars consider the AB sequence intact, i.e. represent $\Delta E' = E_{xz} + E_y - E_{ab}$. We can see that with our expression for ΔE (orange bars), the measured indel position is mostly found near the energy extrema, with short indels (≤ 6 deletions) occupying the high energy percentile, and long indels occupying the low energy percentile. Plots have been gained using a subset of 400 outcomes due to computational constraints.

4.2.4 Conclusion

We have seen that considering the difference in internal energy due to a) torsional stress induced by incorporation of an indel, as well as b) loss of base-stacking interactions over the course of the outcome repair correlates with the observed relative frequency of the respective outcome. This is a smoking gun signal for a promising machine learning algorithm, which will extend our cleavage prediction efforts by the next logical step, i.e. the prediction of the resulting sequence upon repair.

Key takeaways that will inform the full-scale model are therefore as follows:

- Both the energy penalty due to indel-induced effective DNA torsion and the loss of base-stacking interactions due to deletion of a sequence part are justifiable input features for an outcome prediction model.

- NHEJ and (where applicable due to sequence context) MMEJ repair mechanisms should be considered separately.
- External computing power will be needed to generate an appreciable number of data points necessary for the modelling.

4.3 CasTER repair steps modelling

4.3.1 Thermodynamic model of DSB repair

Starting from our considerations above, we developed a simplified thermodynamic model of the two DSB repair pathways which happen most frequently in the absence of an exogenous repair template [152] (see Figure 4.7). The code for this can be found at github.com/florianst/CasTER.

Our energy model for NHEJ is similar to the one in Figure 4.4 and involves the calculation of the net energy difference due to the change in base stacking as

$$E_{\mathrm{NHEJ}} = E_{\mathrm{XZ}} + E_{\mathrm{Y}} - E_{\mathrm{R}} - E_{\mathrm{L}} \tag{4.1}$$

where E_{L} and E_{R} denote the potential energies of the dsDNA to the left and right of the cut site in 3'-5' direction. This constitutes an approximation in the interest of reducing the number of energy calculations, since sequence Y is not necessarily retained as an oligomer but largely resected one base at a time. We chose to not explicitly model the intermediate state $E_{\mathrm{X}} + E_{\mathrm{Y}} + E_{\mathrm{Z}}$ since, given the interactions involved, we do not expect the system to spend an extended amount of time in this state, meaning it does not affect the kinetics of the system.

Analogously, we calculate two energy contributions for the MMEJ repair pathway:

$$E^{(1)}_{\mathrm{MMEJ}} = E_{\mathrm{R}''} + E_{\mathrm{L}'} + E_{\mathrm{r1}} + E_{\mathrm{r2}} + E_{\ell 1} + E_{\ell 2} - E_{\mathrm{R}} - E_{\mathrm{L}}, \tag{4.2}$$

$$E^{(2)}_{\mathrm{MMEJ}} = E_{(\mathrm{LR})''} - E_{\mathrm{R}''} - E_{\mathrm{L}''} \tag{4.3}$$

mirroring the changes in total potential energy due to separation of the strands and resection,

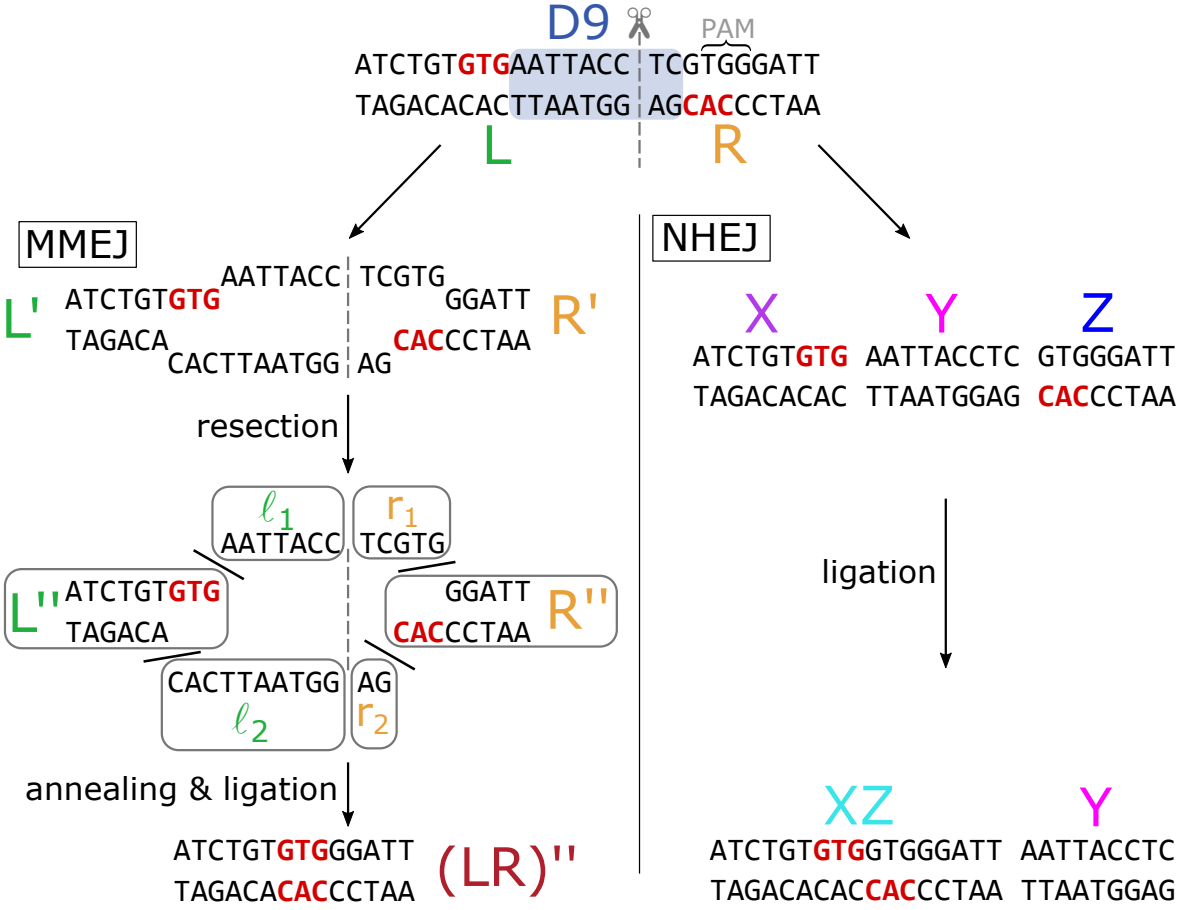


Figure 4.7: Repair step sequences according to the model of non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ) given in [152], here at the example of a 9 bp deletion from the FORECasT dataset [76]. These give rise to our CasTER outcome prediction model.

as well as subsequent annealing and ligation, respectively. Both steps result in a change in base stacking and base pairing interactions.

Additionally, we calculate the torsional contribution as defined in section 4.2

$$E_{\text{torsionDelta}} = E_{\text{strained}} - E_{10.5}. \tag{4.4}$$

In a future version of the model, we envision including a similar torsional contributions for the annealing step during MMEJ as well.

4.3.2 Data source

We again made use of the FORECasT dataset [76], this time using data from two biological replicates (folder *7312067/ST_April_2017_K562_800x_6OA_DPI3_Old7* containing replicate

indel start relative to PAM	indel length 0	1	2	...	20
-4	(0, -4)	(1, -4)	(2, -4)	...	
-5	—	(1, -5)	(2, -5)	...	
-6		—	(2, -6)		
⋮			—	⋱	
					(20, -23)
-24					—

Table 4.1: Visualisation of the 230 possible indel signatures within our chosen ranges for indel length and start position, both measured in base pairs. Each is modelled as one individual label in our multilabel regression model.

A and folder *7312067/ST_Feb_2018_CAS9_12NA_1600X_DPI7* containing replicate B) which were gained 3 and 7 days post transfection, respectively. This will allow for the creation of a diverse training and testing set within the given dataset.

To reduce complexity, we further decided to only look at deletions and not insertions in this study, which covers more than 75% of editing outcomes observed in the FORECasT dataset. It should be a straightforward extension of the model to predict at least 1bp insertions which would bring this figure up to 90%.

4.3.3 Computational data generation

Due to computational constraints, we randomly chose 6000 oligos from biological replicate A and 1000 from biological replicate B and parsed them according to the pattern given in Table 4.1. As a tradeoff between faithfulness to the measured distributions and model tractability, we decided to only include indels up to a maximum length of 20bp as opposed to the limit of 30bp imposed by FORECasT. Despite the long tail of the deletion length distribution, this excludes less than 10% of the total reads for a given oligo.

Once these oligos had been parsed into a custom data structure, we use a batch script to construct a pdb file of a B-DNA of custom sequence as described in section 4.2. This involves temporarily padding the sequence with two A nucleotides either side to make sure

machine type	AWS c6i.metal	Oxford ARC
# vCPUs, type	128 Intel Xeon Ice Lake (3rd Gen)	48 Intel Platinum 8628 (2nd Gen)
CPU frequency	3.5 GHz	2.9 GHz
RAM	256 GB	384 GB
# outcome profiles	6000	1000
# calculations	1,386k	231k
runtime (s)	732k	481k
runtime per calc. (s)	67.6	99.9

Table 4.2: Specifications of IT systems that were used to run the MOSAICS thermodynamics calculations, as well as the observed runtimes. The runtime per calculation was normalised to one vCPU to allow comparison between the second and third generation Intel Xeon processors.

our energy calculation takes into account boundary effects such as end groups, removing hydrogen atoms placed by `fd_helix` and replacing them by hydrogens set using PyMOL, as well as renaming base pairs and hydrogen atoms in the way required by MOSAICS. Given that sequences L″ and R″ have single-stranded overhangs and sequences $\ell_{1,2}$ and $r_{1,2}$ are single-stranded DNA, we remove the appropriate number of nucleotides on the respective strand. Once this is done, MOSAICS is called to calculate the potential energy estimate for the molecule. This was done for all 230 possible indel signatures (see Table 4.1). MMEJ-related energies were only calculated if the respective indel signature resulted in a microhomology at either side of the deleted sequence. Only the longest possible microhomology was considered, given that this should result in the maximum thermodynamic advantage when annealing.

These computations were parallelised and run on high-performance computing hardware. We split this across the Oxford University Advanced Research computing facility and the commercially available Amazon Web Services (AWS) S3 cloud computing facility. A performance and runtime comparison between the two can be found in Table 4.2. Using this combination of computing resources, the total runtime was 8.5 days.

4.3.4 Feature engineering

Energy contributions were calculated according to equations 4.1-4.4 and reshaped into a tabular format with the indel signature as index, leaving NaN for $E^{(1)}_{\text{MMEJ}}$ and $E^{(2)}_{\text{MMEJ}}$ if no microhomology was found. This yields $4 \cdot 230 = 920$ energy features. The potential NHEJ-repaired sequence (XZ in Figure 4.7) was one-hot encoded for each individual indel signature up to a maximum sequence length of 21 bp and appended to the feature columns, yielding another $4 \cdot 21 \cdot 230 = 19,320$ features.

These data points were labelled with the proportion of the total read count for a given oligo that was measured in the FORECasT dataset. Since the authors of FORECasT state that their total read counts for a given oligo are normalised to 1000, data points with a total read count of more than 1000 were ignored. This resulted in a dataset of shape (6804, 20240) with a label array of shape (6804, 230).

4.3.5 Modelling

These data points were split randomly into an 80/20 train/test split and used to train an XGBoost multilabel regressor, where each indel signature corresponds to one label. Hyperparameters were set empirically to `max_depth=3, colsample_bytree=0.7, eta=0.05, n_estimators=200` and a logistic regression loss was used with an RMSE objective. Due to the computational efficiency of the XGBoost gradient boosted tree model implementation, this model can be trained locally in a matter of minutes. We term this model *CasTER* (Cas9 Thermodynamic Energy Repair prediction).

We trained a total of nine versions of this model, adjusting the feature set provided (sequence only, energy only, sequence & energy) as well as the granularity of the prediction through summing over the indel length or position dimension (predicting indel length, position or their joint distribution).

In order to generate feature importance values, we ran a Shapley value explainer on the

trained models using the SHAP library [134]. For computational reasons, this was done on a randomly selected subset of 250 data points within the test set. For each individual point and feature, the library yields a value (*SHAP value*) quantifying its influence on the model output. Summing SHAP values across the feature dimension yields the model prediction for a given point (additivity). In order to quantify the effect of a given feature across data points, we define the *global SHAP value* as the mean absolute SHAP value for that feature.

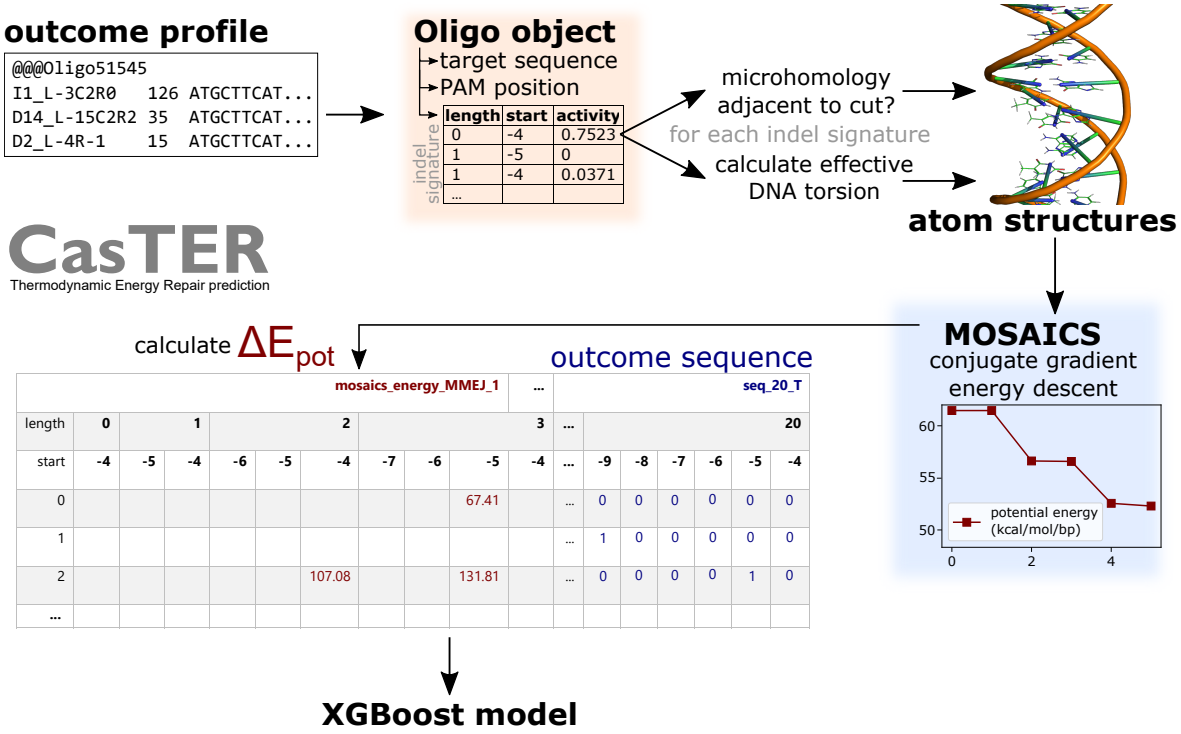


Figure 4.8: Schematic for our CasTER Cas9 DSB repair outcome prediction model. A given outcome profile is parsed into an Oligo object, containing a list of relative read counts for a given indel signature. For each indel signature, we check if it could have been caused by a microhomology as shown in Figure 4.7 and calculate an effective DNA torsion due to strain induced by non-homologous end joining. We build atomic structures representing the pieces of DNA shown in Figure 4.7 and perform a conjugate gradient-based energy minimisation of the structure using MOSAICS [49]. We calculate energy differentials ΔE_{pot} according to equations 4.1 - 4.4. Empty values for ΔE_{pot} signify that no microhomology was found for this particular indel signature. Energy differentials alongside a one-hot representation of the repaired sequence are assembled into a dataframe, which is used to train a boosted tree regression model that predicts relative read counts for each individual indel signature.

4.3.6 Results & Discussion

Table 4.3 shows a wide range of benchmark values that quantify the deviation between true and predicted outcome distributions, using both our model and the FORECasT command line model which can be downloaded online. In order to achieve comparability with Allen et al. [76], we use the symmetrised Kullback–Leibler divergence KL as one performance benchmark:

$$\mathrm{KL} = D_{\mathrm{KL}}(P_{\mathrm{true}}||P_{\mathrm{pred}}) + D_{\mathrm{KL}}(P_{\mathrm{true}}||P_{\mathrm{true}}), \tag{4.5}$$

$$D_{\mathrm{KL}}(P||Q) = \sum_{i \in \mathcal{S}} P_i \log \frac{P_i}{Q_i} \tag{4.6}$$

where $\mathcal{S}$ represents the set of all considered indel signatures as defined in Table 4.1. A small pseudocount of 0.5 is added to indel signatures which are present in one distribution but not the other in order to avoid division by zero. This is consistent with the approach in [76].

Table 4.3 shows that our CasTER model outperforms FORECasT in direct comparison for most configurations. Exceptions are according to Spearman correlation and Kendall tau when predicting indel position and the joint indel length and position distribution. We further see that adding energy features improves CasTER performance as compared to pure sequence-based prediction when considering KL divergence between true and predicted outcome distribution. This supports our hypothesis that energy features as described above can inform predictor models more than purely sequence-based features.

We note that the Pearson correlation between true and predicted indel length as well as joint indel length and position distributions is an exception to this trend, with this benchmark value either slightly decreasing or changing negligibly when introducing energy features. We attribute this to the strongly bimodal distribution of the CasTER prediction (see Figure 4.9) as compared with FORECasT predictions, which makes it less suitable for analysis via established correlation quantifiers. It should be noted that KL divergence is the only benchmark given here that considers a given outcome distribution as a whole rather than a flattened version of it, making it a more realistic benchmark. Figure 4.10 shows that the sym-

predicting	model	feature set	KL div	Pears	Spear	K-τ	R^2
indel length	CasTER	sequence & energy	0.284	0.945	**0.376**	**0.273**	0.892
		energy only	**0.282**	0.943	0.366	0.265	0.890
		sequence only	0.308	**0.947**	0.330	0.238	**0.897**
	FORECasT	sequence only	0.722	0.880	0.364	0.266	0.774
indel position	CasTER	sequence & energy	**0.261**	**0.959**	0.567	0.432	**0.919**
		energy only	**0.260**	**0.959**	0.565	0.430	**0.920**
		sequence only	0.268	0.958	0.549	0.417	0.917
	FORECasT	sequence only	0.531	0.902	**0.615**	**0.479**	0.814
indel length and position	CasTER	sequence & energy	**0.274**	**0.951**	0.295	0.236	**0.905**
		energy only	**0.274**	0.948	0.285	0.228	0.898
		sequence only	0.285	**0.952**	0.252	0.201	**0.906**
	FORECasT	sequence only	0.696	0.888	**0.341**	**0.310**	0.789
	inDelphi*	sequence only	—	0.840	—	—	—
	between replicates	sequence only	0.657	0.918	0.132	0.126	0.842

Table 4.3: Benchmark values when using our model CasTER and the FORECasT model [76] on our chosen portion of the FORECasT data set. A literature comparison using the inDelphi model [77] is denoted by an asterisk. Our chosen benchmarks include the Kullback-Leibler divergence between the true and predicted distribution, Pearson, Spearman and Kendall tau correlations between the sets of true and predicted activity labels, as well as the coefficient of determination as the square of the Pearson correlation which is given here for ease of literature comparison. Benchmarks have been calculated for separate models predicting indel length and position as well as predicting their joint distribution. The last line in the table gives the respective benchmarks when applied not to true and predicted activity labels but to the two biological replicates in the FORECasT dataset.

metrised KL divergence distribution which we obtained running FORECasT on our chosen subset of data lines up well in average and width with the values given in the FORECasT publication.

Figure 4.9 further illustrates the strong predictive power of our model which separates high from low activity indel signatures with high accuracy. The pronounced high activity mode is largely due to our model predicting perfect repair with an average of 78% in the test set, which is essentially identical to the average ground truth value, whereas for the FORECasT predictions this mode seems to be centred around 50%.

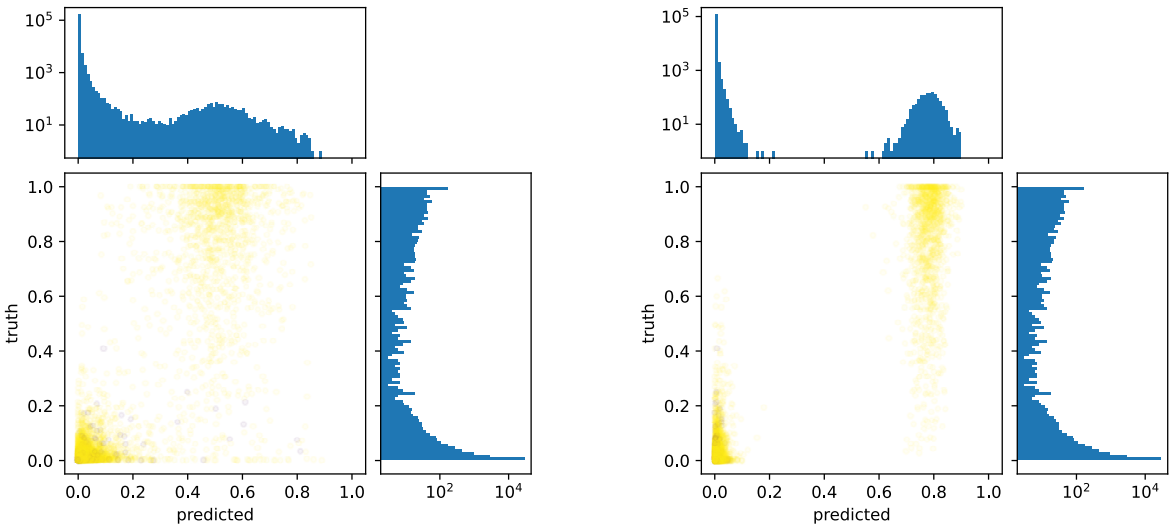


Figure 4.9: Scatter plot of the true vs. predicted activity label of a given indel signature within our chosen portion of the FORECasT data set [76]. The left plot has been generated by running the FORECasT model, the right plot using our CasTER model predicting the joint indel length and position distribution from both sequence and energy features. This visualisation includes all indel signatures. Around 9% of data points had a sequence context which did not allow for MMEJ; these are coloured darker. Histogram projections onto the true and predicted label axes are shown for reference (note the logarithmic scale).

Figure 4.13 shows the SHAP summary plot for the model predicting the joint indel length and position distribution from both sequence and energy features. This can be considered the central figure of this chapter, as it contains a wealth of information about the relative predictive power of energy and sequence features, as well as the nature of their influence on

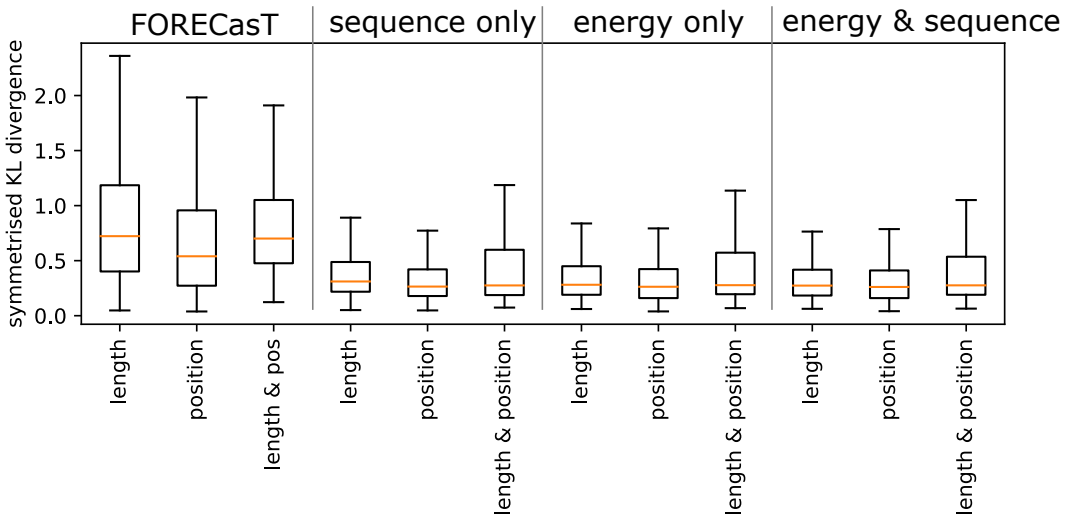


Figure 4.10: Distributions of symmetrised Kullback-Leibler (KL) divergence between predicted and true outcome distribution. The median of the symmetrised KL distributions is quoted in Table 4.3. Predictions were gained using FORECasT [76] (left three boxes) and our models. The latter incorporate either only sequence features, only energy features or both. For all models, boxes for predicting the distribution of indel length and indel position separately, as well as their joint distribution are shown.

model prediction. Firstly, we observe that there is no sequence-based feature among the top 15 most influential features as measured by global SHAP value, and only three energy features among the top 35 most influential features. This underlines the power and relevance of our fundamentally novel, structure-based approach. It is the use of this novel class of energy features, introducing domain knowledge about atomic forces governing the interaction between nucleotides, combined with a powerful gradient-boosted tree model that sets it apart from existing approaches such as FORECasT [76]. Especially the energy difference $E^{(2)}_{\text{MMEJ}}$ through annealing and ligating the microhomologous regions is shown to be a highly influential feature class. Secondly, we observe that the influence of all shown energy contributions is of the shape we would expect from DSB repair as a probabilistic, energy-driven process, with energetically favourable feature values showing positive SHAP values, i.e. an increased predicted activity, and vice versa.

In order to verify that the considerable importance of some of the $E^{(2)}_{\text{MMEJ}}$ features is not

a direct cause of the length of the annealing microhomology, we look at the SHAP value of these features by microhomology length. Figure 4.11 shows no discernable difference in the underlying distributions of SHAP values for any of the microhomology classes.

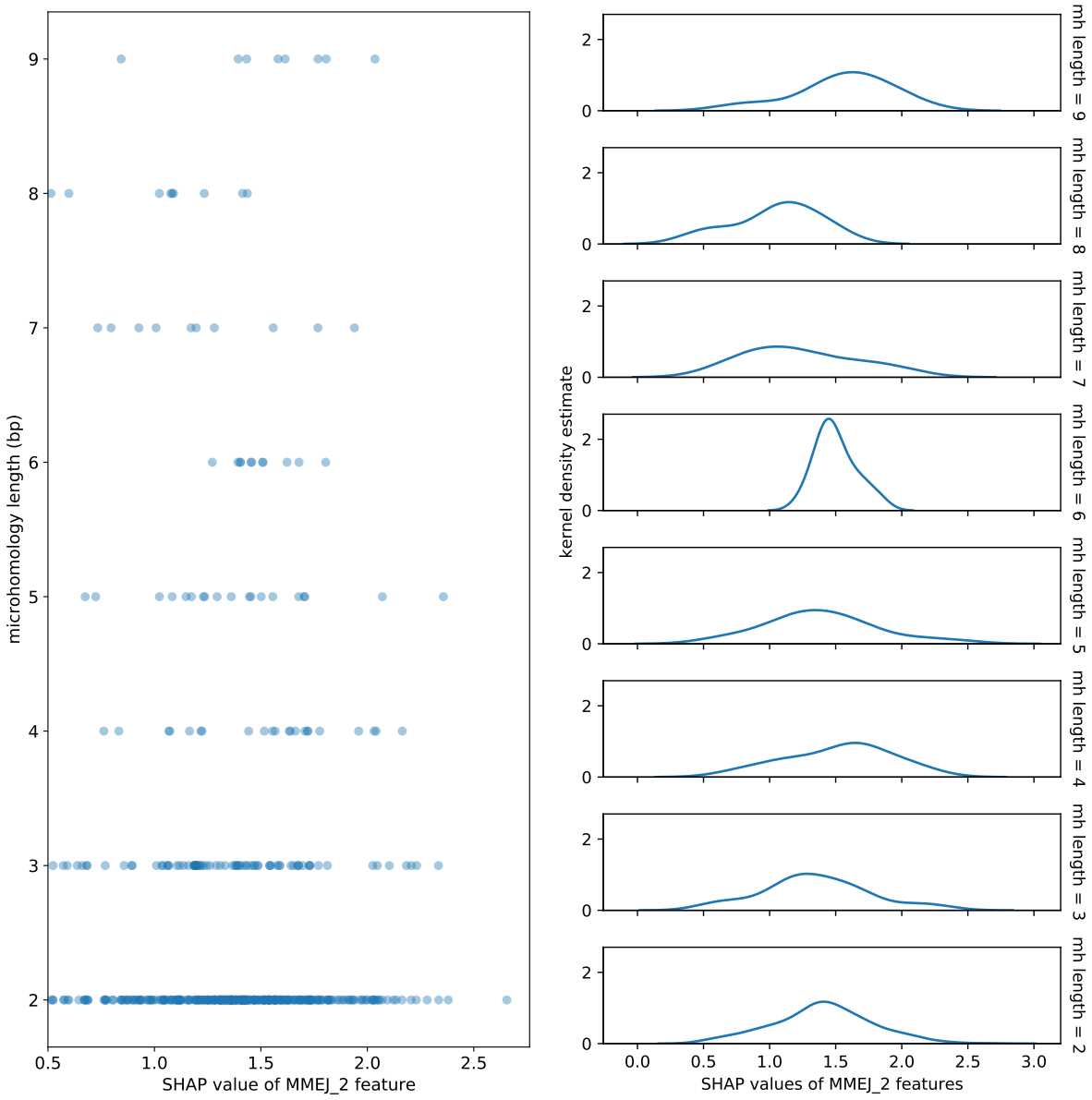


Figure 4.11: SHAP values of the $E^{(2)}_{\text{MMEJ}}$ features by length of the annealing microhomology. The right panel shows a kerndel density estimate of the underlying distribution, which appears identical up to statistical variations due to the low number of long homologies in the SHAP value dataset.

Allen *et al.* propose three main sequence determinants that favour reproducible deletions [76], namely the presence of microhomologies either side of the cut site, short distances between microhomologies and repeating nucleotides at the cut site. While we can confirm the

first two notions with our trained model using Figures 4.12 and 4.13, the third is not unambiguously verifiable in our current model version without including dinucleotide features at the cut site into the model explanation. We plan to include these in a future version of the model, alongside with other hand-crafted features such as repeat length at the cut site, based on their relevance in [76].

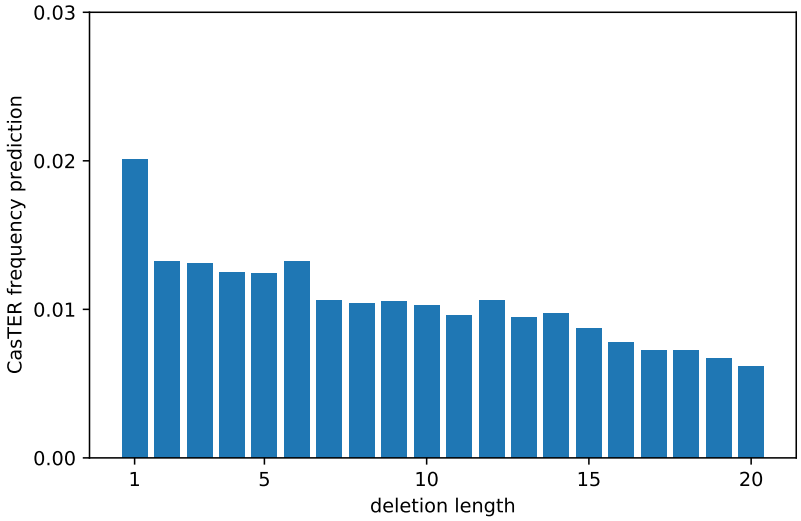


Figure 4.12: Histogram of CasTER repair outcome frequency predictions by predicted deletion length, averaged over all MMEJ-related data points within the test set. Deletion lengths of 0 (perfect repair) are not shown.

In order to quantify how the trained model attributes importances to energy feature values, Figure 4.14 shows a SHAP dependence plot of all the $E^{(2)}_{\mathrm{MMEJ}}$ features combined. In order to achieve comparability of the shapes of individual distributions for this set of features, we align two percentile values for each individual $E^{(2)}_{\mathrm{MMEJ}}$ input feature across all indel signatures. We hypothesise that if the model draws on features in a thermodynamically inspired way, the set of all SHAP values aligned in this way should approach a negative exponential (Boltzmann) distribution. We indeed observe a negative exponential shape, except for some of the MMEJ energy values between 50 and 100 kcal/mol. Here, the model attributes high importance to these features, hence the SHAP value frequently lies above its 95th percentile. In this range, the model seems to discriminate less well between energy values, leading to

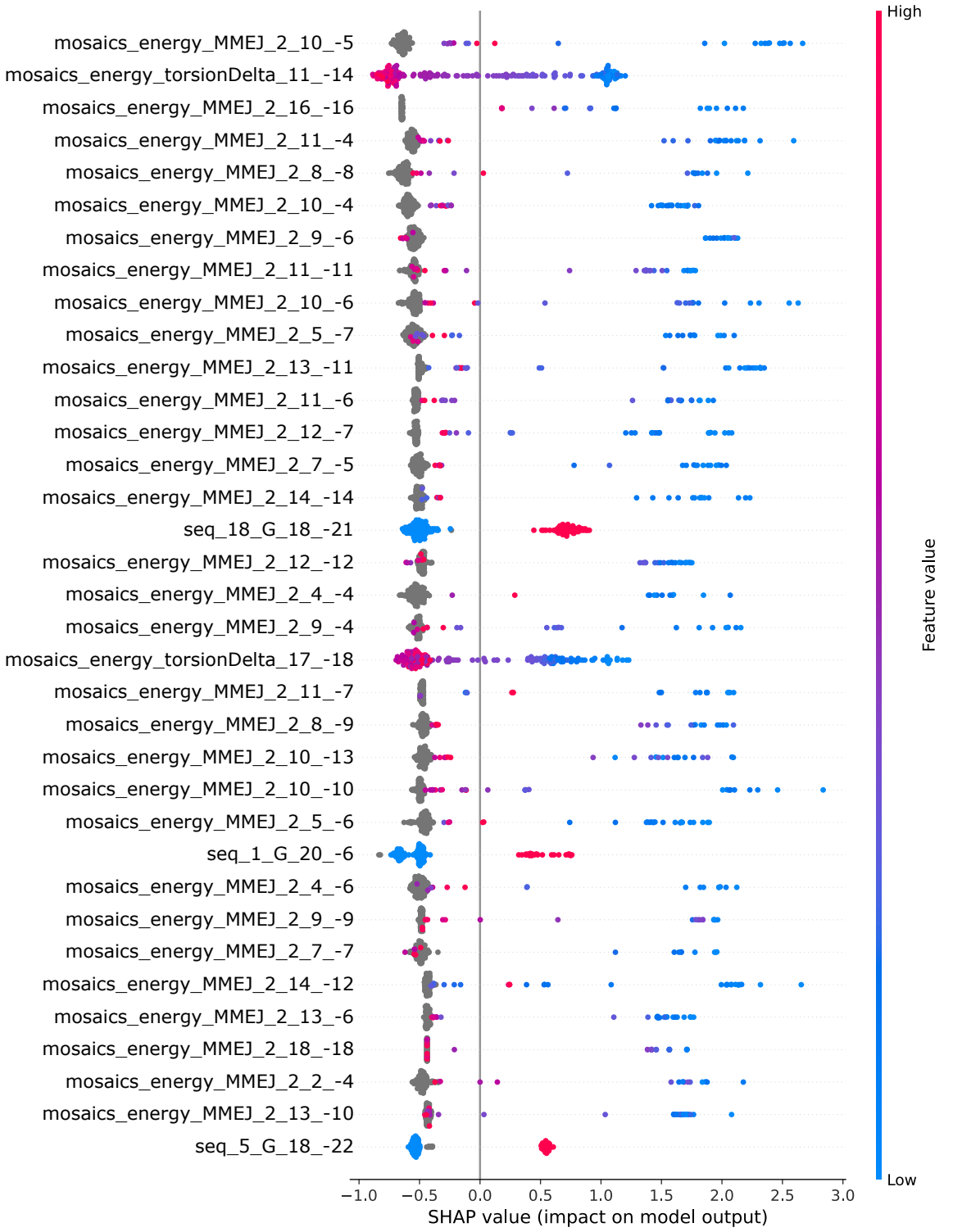


Figure 4.13: SHAP feature importance values for the energy + sequence model. Features which have not been present for a particular data point are coloured in gray. We observe that energetically favourable (low / blue) values of torsionDelta result in increased predicted activity (positive SHAP value). Likewise, for data points with microhomologies, energetically more favourable microhomology-mediated end joining MMEJ energy differences (low / blue) result in an increased predicted activity, with the lowest values (bright blue) often resulting in the largest increase. The absence of a microhomology for the respective indel signature (gray points) has a similar effect on the model as a present but energetically unfavourable microhomology (red points).

them bunching up. For energy values above 100 however, we get a good agreement with a negative exponential shape, indicating that the model indeed learns a thermodynamic behaviour.

Finally, when averaging over data points and indel signatures, we can extract the influence of certain nucleotides at certain positions within the target sequence on the predicted indel frequency. This is shown in Figure 4.15. Due to the strong bimodality of the distribution of SHAP values for sequence features, with low feature values (nucleotide not set) showing negative SHAP values (see Figure 4.13), we here show the global SHAP value as the difference between the mean SHAP value for 1 (nucleotide set) and the mean SHAP value for 0 (nucleotide not set) entries in the one-hot encoding. We observe a comparatively strong positive influence of G nucleotides in positions 3 to 10 of the target sequence, i.e. the nucleotides preceding the cut site by 8 to 15 bp. To our knowledge, this effect has not been reported in literature yet. We note that the centre of this region is around 11 bp away from the cut site which lines up with the 10.5 bp helical periodicity of the DNA. This supports our initial hypothesis about the connection between this periodicity and the outcome distribution of DSB repair.

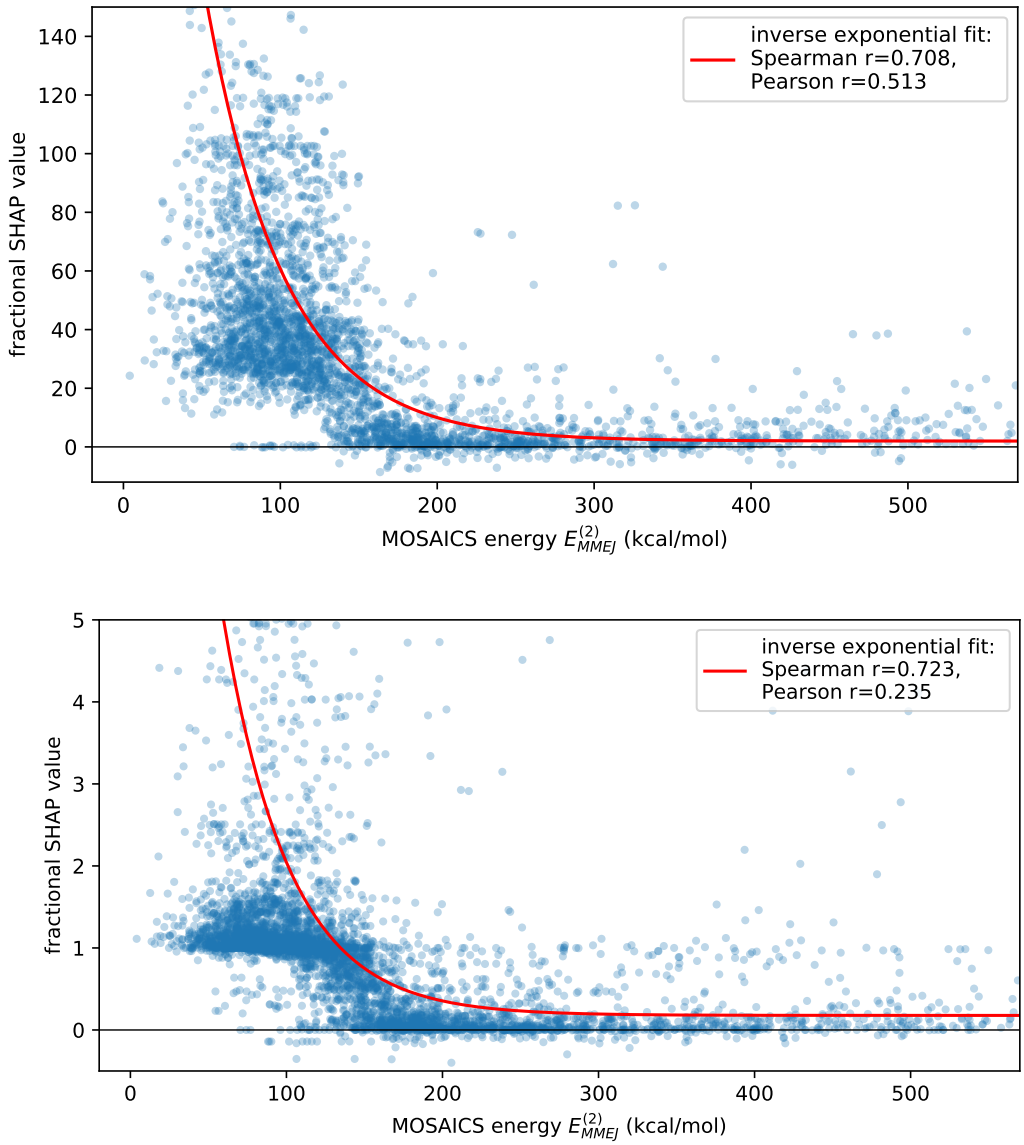
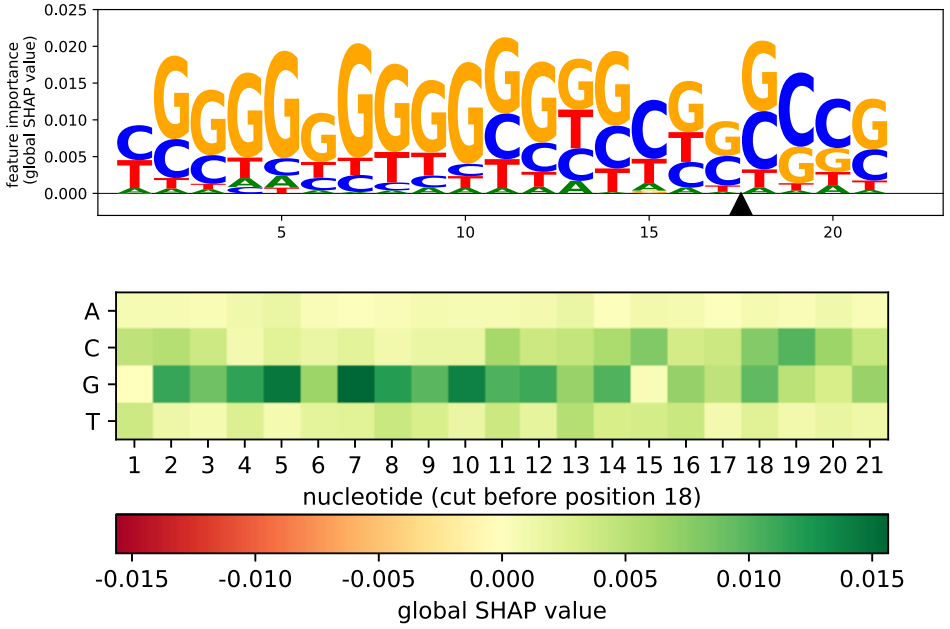


Figure 4.14: SHAP dependence plot of all the $E^{(2)}_{\text{MMEJ}}$ features combined. We obtained the fractional SHAP value by aligning the 5th and 70th/95th percentile (top/bottom panel) of SHAP values. As indicated by the negative exponential fit line shown in red, these two quantities approach a Boltzmann distribution with an effective $k_{\text{B}}^{\text{eff}}T \approx 0.0225$ kcal/mol (top panel), indicating that the model learns a thermodynamic dependence between $E^{(2)}_{\text{MMEJ}}$ and the likelihood of a certain indel outcome. The fit and SHAP values have a Spearman correlation of $r = 0.708$ (top panel). As in Figure 4.13, we combine the SHAP values for all indel signatures here and use the energy + sequence model.

A



B

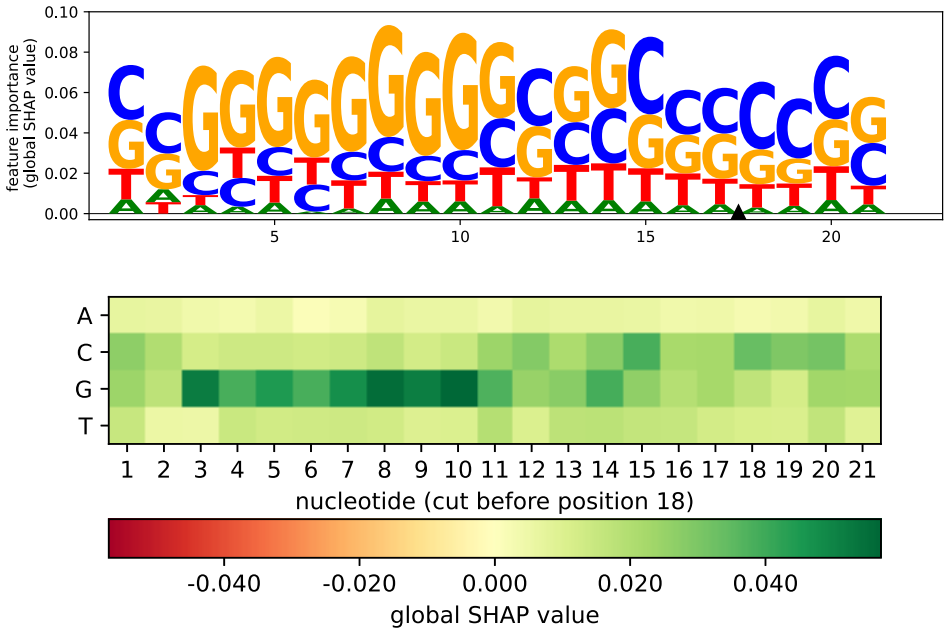


Figure 4.15: Sequence logo (top row) and heatmap (bottom row) representations of the global SHAP feature importance value for the target sequence. This is shown for the sequence + energy model (**A**) and the sequence-only model (**B**).

Chapter 5

Conclusion & Further Developments

In **chapter 2**, we have developed and implemented a bioinformatics pipeline that can aggregate diverse data sources quantifying CRISPR/Cas off-target cleavage into a single dataset. It can further annotate this data with appropriate features relevant to off-target cleavage modelling. A user-friendly web interface was created, offering usability for a wide range of users.

We envision to extend the database by appropriately annotated studies targeting high-fidelity Cas9 nuclease variants and nucleases from different organisms, as well as Cas9 off-target studies on different vertebrate organisms.

In order to provide a one-stop experience for off-target effects, we propose the inclusion of a state-of-the-art cleavage prediction algorithm which can provide predicted off-target effects next to measured off-target effects for a given gRNA.

Chapter 3 then concerned the development of an off-target cleavage predictor. Through the introduction of a new feature class and the careful adjustment of model architectures, we have identified a set of models which match the performance of state-of-the-art off-target cleavage prediction algorithms in direct comparison. All models are highly influenced by nucleosome organisation-related features such as histone binding affinity, which emphasises the importance of capturing the biological environment around the cleavage site when modelling cleavage activity. Our approach has shown that these computed physically informed

features are fit to enhance the predictive power of cleavage prediction models and to replace experimental epigenetic markers in future modelling efforts. We have further provided an accessible, user-friendly command line interface that allows users of various disciplines to utilise all our models, even without providing a complete set of features. This all paves the way towards the prediction of off-target sites which would so far have gone unnoticed.

Our environmentally sensitive set of features reveals several novel, promising pathways towards further improvement of off-target cleavage prediction. Going forward, it could be fruitful to increase model complexity, e.g. using a 2D convolutional kernel to capture interaction between features of a single nucleotide. A 2D convolution kernel would be able to capture the base-pair resolved interaction between sequence and nucleosomal markers as well as between sequence k-mers. Further than this, our multimodal data could be fused at different stages of the model, such that sequence, nucleosomal and energy features interact at differing levels of representation (*data integration*).

We further envision to replace the epigenetic information of the guide, which so far only copies the epigenetic information of the target DNA. This is clearly an unphysical choice. Given that a synthetic sgRNA does by design not carry epigenetic markers, a one-hot encoded dot-bracket representation of the sgRNA spacer secondary structure would be a more suitable choice to capture its informative properties. As argued in reference [153], sgRNA-internal hairpins can lead to either the inability of such gRNAs to form a functional Cas9-gRNA complex or their inability to recognise targets, with both phenomena depending mainly on the structure of the spacer part.

Through the implementation, benchmarking and interpretation of a DSB repair outcome distribution predictor in **chapter 4**, we have demonstrated that thermodynamic features based on structural DNA calculations carry valuable information about the process of DSB repair. It is the inclusion of this feature class simultaneously for the entire set of indel signatures, especially those features pertaining to NHEJ-related torsion and MMEJ-related microhomology

annealing, that set apart our algorithm from existing approaches. Using appropriate computational power, these features can be readily generated for arbitrary cleavage distributions. Interacting with the MOSAICS software package used to generate the thermodynamic features on a closer level than its command line interface would likely even accelerate this process and reduce the required processing time.

Our trained DSB repair prediction model CasTER can directly use cleavage distributions generated by our off-target cleavage prediction model piCRISPR presented in the previous chapter. Together, they provide a one-stop, end-to-end computational source for a comprehensive overview of the effects of CRISPR/Cas9 gene editing with a given sgRNA. Experts can then judge the predicted effects of the chosen sgRNA on the genome with respect to potential safety concerns. With an appropriate dataset of safety-annotated gene regions, a software tool would even be able to aid in this task, providing a safety score for a given sgRNA. This would allow for an automatised suggestion of ideal sgRNAs to target a specific DNA region or locus whilst minimising known risks for the organism.

In this sense, we have provided computational insights, methods and software tools that are fit to enhance the safety of CRISPR/Cas9 gene editing.

Bibliography

[1] *Human Genome Editing: Science, Ethics, and Governance*. The National Academies Press, Washington, DC 2017. ISBN: 978-0-309-45288-5.

[2] Glass *et al.*, *Focus on Applications of CRISPR Engineering the Delivery System for CRISPR-Based Genome Editing*. Trends Biotechnol. **36**, 173 (2017).

[3] Kajaste-Rudnitski *et al.*, *Cellular innate immunity and restriction of viral infection: Implications for lentiviral gene therapy in human hematopoietic cells*. Hum. Gene Ther. **26**, 201 (2015).

[4] Davies, *The technical risks of human gene editing*. Hum. Reprod. **34**, 2104 (2019).

[5] Kosicki *et al.*, *Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements*. Nat. Biotechnol. **36**, 765 (2018).

[6] Alanis-Lobato *et al.*, *Frequent loss of heterozygosity in CRISPR-Cas9-edited early human embryos*. Proc. Natl. Acad. Sci. U. S. A. **118**, e2004832117 (2021).

[7] Anliker *et al.*, *Regulatory Considerations for Clinical Trial Applications with CRISPR-Based Medicinal Products*. Cris. J. **5**, 364 (2022).

[8] Dai *et al.*, *CRISPR-Cas9 for in vivo Gene Therapy: Promise and Hurdles*. Mol. Ther. - Nucleic Acids **5**, e349 (2016).

[9] Fellmann *et al.*, *Cornerstones of CRISPR-Cas in drug discovery and therapy*. Nat. Rev. Drug Discov. **16**, 89 (2017).

[10] Sun *et al.*, *Advances in therapeutic application of CRISPR-Cas9*. Brief. Funct. Genomics **19**, 164 (2020).

[11] Buechele *et al.*, *MLL leukemia induction by genome editing of human CD34+ hematopoietic cells*. Blood **126**, 1683 (2015).

[12] Farasat *et al.*, *A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation*. PLoS Comput. Biol. **12**, e1004724 (2016).

[13] Alkan *et al.*, *CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters*. Genome Biol. **19**, 177 (2018).

[14] Ishino *et al.*, *Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.* J. Bacteriol. **169**, 5429 (1987).

[15] Horvath *et al.*, *CRISPR/Cas, the immune system of bacteria and archaea*. Science **327**, 167 (2010).

[16] Wang *et al.*, *CRISPR/Cas9 in Genome Editing and Beyond*. Annu. Rev. Biochem. **85**, 227 (2016).

[17] Kunin *et al.*, *Evolutionary conservation of sequence and secondary structures in CRISPR repeats*. Genome Biol. **8**, R61 (2007).

[18] Urnov *et al.*, *Genome editing with engineered zinc finger nucleases*. Nat. Rev. Genet. **11**, 636 (2010).

[19] Joung *et al.*, *TALENs: a widely applicable technology for targeted genome editing*. Nat. Rev. Mol. Cell Biol. **14**, 49 (2013).

[20] Doudna, *CRISPR Systems - Doudna Lab*. https//doudnalab.org/research_areas/ crispr-systems (2020).

[21] Kim *et al.*, *Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells*. Nat. Methods **12**, 237 (2015).

[22] Anders *et al.*, *Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease*. Nature **513**, 569 (2014).

[23] Jinek *et al.*, *A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity*. Science **337**, 816 (2012).

[24] Jinek *et al.*, *Structures of Cas9 endonucleases reveal RNA-mediated conformational activation*. Science **343**, 1247997 (2014).

[25] Ozaki *et al.*, *Role of p53 in Cell Death and Human Cancers*. Cancers **3**, 994 (2011).

[26] Nekrasov *et al.*, *Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion*. Sci. Rep. **7**, 482 (2017).

[27] Tsai *et al.*, *GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases*. Nat. Biotechnol. **33**, 187 (2015).

[28] Hu *et al.*, *Detecting DNA double-stranded breaks in mammalian genomes by linear amplification–mediated high-throughput genome-wide translocation sequencing*. Nat. Protoc. **11**, 853 (2016).

[29] Cameron *et al.*, *Mapping the genomic landscape of CRISPR–Cas9 cleavage*. Nat. Methods **14**, 600 (2017).

[30] Tsai *et al.*, *CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets*. Nat. Methods **14**, 607 (2017).

[31] Lazzarotto *et al.*, *CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity*. Nat. Biotechnol. **38**, 1317 (2020).

[32] Kwon *et al.*, *Extru-seq: a method for predicting genome-wide Cas9 off-target sites with advantages of both cell-based and in vitro approaches*. Genome Biol. **24(4)**, 1 (2023).

[33] Manghwar *et al.*, *CRISPR/Cas Systems in Genome Editing: Methodologies and Tools for sgRNA Design, Off-Target Evaluation, and Strategies to Mitigate Off-Target Effects*. Adv. Sci. **7**, 1902312 (2020).

[34] Ran *et al.*, *In vivo genome editing using Staphylococcus aureus Cas9*. Nature **520**, 186 (2015).

[35] Zetsche *et al.*, *Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System*. Cell **163**, 759 (2015).

[36] Abudayyeh *et al.*, *C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector*. Science **353** (2016).

[37] Kleinstiver *et al.*, *High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects*. Nature **529**, 490 (2016).

[38] Slaymaker *et al.*, *Rationally engineered Cas9 nucleases with improved specificity*. Science **351**, 84 (2016).

[39] Truong *et al.*, *Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells*. Proc. Natl. Acad. Sci. U. S. A. **110**, 7720 (2013).

[40] ABM Inc., *CRISPR Cas9*. old.abmgood.com/marketing/knowledge_base/CRISPR_Cas9_Introduction.php (2020).

[41] Levitt, *How many base-pairs per turn does DNA have in solution and in chromatin? Some theoretical calculations.* Proc. Natl. Acad. Sci. U. S. A. **75**, 640 (1978).

[42] SantaLucia *et al.*, *The thermodynamics of DNA structural motifs*. Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).

[43] Mak, *Unraveling base stacking driving forces in DNA*. J. Phys. Chem. B **120**, 6010 (2016).

[44] Durrant *et al.*, *Molecular dynamics simulations and drug discovery*. BMC Biol. **9**, 71 (2011).

[45] Pérez *et al.*, *Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers*. Biophys. J. **92**, 3817 (2007).

[46] Murphy, *Machine Learning*. 2012. ISBN: 9780262018029.

[47] Chen *et al.*, "XGBoost: A scalable tree boosting system". In: *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 2016.

[48] Goodfellow *et al.*, *Deep Learning*. MIT Press, 2017.

[49] Sim *et al.*, *Modeling and design by hierarchical natural moves*. Proc. Natl. Acad. Sci. **109**, 2890 (2012).

[50] Baryshnikova, *Data libraries – the missing element for modeling biological systems*. FEBS J. **287**, 4594 (2020).

[51] Rousseau *et al.*, *CRISPI: a CRISPR interactive database*. Bioinformatics **25**, 3317 (2009).

[52] Hodgkins *et al.*, *WGE: a CRISPR database for genome engineering*. Bioinformatics **31**, 3078 (2015).

[53] Jaskula-Ranga *et al.*, *grID: A CRISPR-Cas9 guide RNA Database and Resource for Genome-Editing*. bioRxiv, 097352 (2016).

[54] Rauscher *et al.*, *GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens*. Nucleic Acids Res. **45**, D679 (2017).

[55] Edgar *et al.*, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository*. Nucleic Acids Res. **30**, 207 (2002).

[56] ENCODE Project Consortium, *The ENCODE (ENCyclopedia Of DNA Elements) Project*. Science **306**, 636 (2004).

[57] Barrett *et al.*, *BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata*. Nucleic Acids Res. **40**, D57 (2012).

[58] Weng, *SCREEN: Search Candidate cis-Regulatory Elements by ENCODE*. http://screen.encodeproject.org/ (2017).

[59] Oughtred *et al.*, *The BioGRID interaction database: 2019 update*. Nucleic Acids Res. **47**, 529 (2018).

[60] Listgarten *et al.*, *Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs*. Nat. Biomed. Eng. **2**, 38 (2018).

[61] Liu *et al.*, *Computational approaches for effective CRISPR guide RNA design and evaluation*. Comput. Struct. Biotechnol. J. **18**, 35 (2020).

[62] Dimauro *et al.*, *A systematic mapping study on machine learning techniques for the prediction of CRISPR / Cas9 sgRNA target cleavage*. Comput. Struct. Biotechnol. J. **20**, 5813 (2022).

[63] Liu *et al.*, *Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas System using attention boosted deep learning and network-based gene feature*. PLoS Comput. Biol. **15**, e1007480 (2019).

[64] Charlier *et al.*, *Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-Cas9 gene editing*. Bioinformatics **37**, 2299 (2021).

[65] Liu *et al.*, *Deep learning improves the ability of sgRNA off-target propensity prediction*. BMC Bioinformatics **21**, 51 (2020).

[66] Lin *et al.*, *Off-target predictions in CRISPR-Cas9 gene editing using deep learning*. Bioinformatics **34**, 656 (2018).

[67] Lin *et al.*, *CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels*. Adv. Sci. **7**, 1903562 (2020).

[68] Zhang *et al.*, *Effective use of sequence information to predict CRISPR-Cas9 off-target*. Comput. Struct. Biotechnol. J. **20**, 650 (2022).

[69] Fu *et al.*, *Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity*. Nat. Commun. **13**, 474 (2022).

[70] Guan *et al.*, *Transformer-based anti-noise models for CRISPR-Cas9 off-target activities prediction*. Brief. Bioinform. **24(3)**, 1 (2023).

[71] Li *et al.*, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics **25**, 1754 (2009).

[72] Langmead *et al.*, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biol. **10**, R25 (2009).

[73] Bae *et al.*, *Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases*. Bioinformatics **30**, 1473 (2014).

[74] Canzar *et al.*, *Short Read Mapping: An Algorithmic Tour*. Proc. IEEE **105**, 436 (2017).

[75] Chuai *et al.*, *DeepCRISPR: Optimized CRISPR guide RNA design by deep learning*. Genome Biol. **19**, 80 (2018).

[76] Allen *et al.*, *Predicting the mutations generated by repair of Cas9-induced double-strand breaks*. Nat. Biotechnol. **37**, 64 (2018).

[77] Shen *et al.*, *Predictable and precise template-free CRISPR editing of pathogenic variants*. Nature **563**, 646 (2018).

[78] Chen *et al.*, *Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair*. Nucleic Acids Res. **47**, 7989 (2019).

[79] Leenay *et al.*, *Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells*. Nat. Biotechnol. **37**, 1034 (2019).

[80] Liu *et al.*, *Predicting CRISPR/Cas9 Repair Outcomes by Attention-Based Deep Learning Framework*. Cells **11**, 1847 (2022).

[81] Mathis *et al.*, *Predicting prime editing efficiency and product purity by deep learning*. Nat. Biotechnol. doi.org/10.1038/s41587–022–01613 (2023).

[82] Arbab *et al.*, *Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning*. Cell **182**, 463 (2020).

[83] Störtz *et al.*, *crisprSQL: a novel database platform for CRISPR/Cas off-target cleavage assays*. Nucleic Acids Res. **49**, 855 (2020).

[84] Han *et al.*, *Mitigating off-target effects in CRISPR/Cas9-mediated in vivo gene editing*. J. Mol. Med. **98**, 615 (2020).

[85] Vakulskas *et al.*, *Evaluation and Reduction of CRISPR Off-Target Cleavage Events*. Nucleic Acid Ther. **29**, 167 (2019).

[86] Gao *et al.*, *Data imbalance in CRISPR off-target prediction*. Brief. Bioinform. **21**, 1448 (2020).

[87] Dunham *et al.*, *An integrated encyclopedia of DNA elements in the human genome*. Nature **489**, 57 (2012).

[88] Liu *et al.*, *Multi-omic measurements of heterogeneity in HeLa cells across laboratories*. Nat. Biotechnol. **37**, 314 (2019).

[89] Frattini *et al.*, *High variability of genomic instability and gene expression profiling in different HeLa clones*. Sci. Rep. **5**, 15377 (2015).

[90] Fu *et al.*, *High frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells*. Nat. Biotechnol. **31**, 822 (2013).

[91] Cho *et al.*, *Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases*. Genome Res. **24**, 132 (2014).

[92] Ran *et al.*, *In vivo genome editing using Staphylococcus aureus Cas9*. Nature **520**, 186 (2015).

[93] Frock *et al.*, *Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases*. Nat. Biotechnol. **33**, 179 (2015).

[94] Wang *et al.*, *Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors*. Nat. Biotechnol. **33**, 175 (2015).

[95] Kim *et al.*, *Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq*. Genome Res. **26**, 406 (2016).

[96] Chen *et al.*, *Enhanced proofreading governs CRISPR–Cas9 targeting accuracy*. Nature **550**, 407 (2017).

[97] Kim *et al.*, *DIG-seq: a genome-wide CRISPR off-target profiling method using chromatin DNA*. Genome Res. **28**, 1894 (2018).

[98] Anderson *et al.*, *CRISPR off-target analysis in genetically engineered rats and mice*. Nat. Methods **15**, 512 (2018).

[99] Jones *et al.*, *Massively parallel kinetic profiling of natural and engineered CRISPR nucleases*. Nat. Biotechnol. **39**, 84 (2021).

[100] Haeussler *et al.*, *Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR*. Genome Biol. **17**, 148 (2016).

[101] Hsu *et al.*, *DNA targeting specificity of RNA-guided Cas9 nucleases*. Nat. Biotechnol. **31**, 827 (2013).

[102] Newman *et al.*, *Cas9 Cuts and Consequences; Detecting, Predicting, and Mitigating CRISPR/Cas9 On- and Off-Target Damage*. BioEssays **2020**, 2000047 (2020).

[103] Zuo *et al.*, *Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations*. Sci. Rep. **6**, 37584 (2016).

[104] Atkins *et al.*, *Off-Target Analysis in Gene Editing and Applications for Clinical Translation of CRISPR/Cas9 in HIV-1 Therapy*. Front. Genome Ed. **3**, 673022 (2021).

[105] O'Geen *et al.*, *How specific is CRISPR/Cas9 really?* Curr. Opin. Chem. Biol. **29**, 72 (2015).

[106] Frankish *et al.*, *GENCODE reference annotation for the human and mouse genomes*. Nucleic Acids Res. **47**, D766 (2019).

[107] Moore *et al.*, *Expanded encyclopaedias of DNA elements in the human and mouse genomes*. Nature **583**, 699 (2020).

[108] Zhang *et al.*, *C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks*. Comput. Struct. Biotechnol. J. **18**, 344 (2020).

[109] Franco *et al.*, *The Role of CCCTC-Binding Factor (CTCF) in Genomic Imprinting, Development, and Reproduction*. Biol. Reprod. **91** (2014).

[110] Sims *et al.*, *Histone lysine methylation: A signature for chromatin function*. Trends Genet. **19**, 629 (2003).

[111] De Magis *et al.*, *DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells.* Proc. Natl. Acad. Sci. U. S. A. **116**, 816 (2019).

[112] Neph *et al.*, *BEDOPS: high-performance genomic feature operations*. Bioinformatics **28**, 1919 (2012).

[113] Nadel *et al.*, *RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships*. Epigenetics and Chromatin **8**, 46 (2015).

[114] Sanz *et al.*, *Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals*. Mol. Cell **63**, 167 (2016).

[115] Gorthi *et al.*, *EWS-FLI1 increases transcription to cause R-Loops and block BRCA1 repair in Ewing sarcoma*. Nature **555**, 387 (2018).

[116] Barrett *et al.*, *NCBI GEO: archive for functional genomics data sets-update*. Nucleic Acids Res. **41**, D991 (2013).

[117] Barnett *et al.*, *BamTools: a C++ API and toolkit for analyzing and managing BAM files*. Bioinformatics **27**, 1691 (2011).

[118] Gruber *et al.*, *The Vienna RNA Websuite*. Nucleic Acids Res. **36**, W70 (2008).

[119] Wang *et al.*, *CRISPR-GNL: an improved model for predicting CRISPR activity by machine learning and featurization*. bioRxiv **605790** (2019).

[120] Box *et al.*, *An Analysis of Transformations*. J. R. Stat. Soc. B **26**, 211 (1964).

[121] Yuan *et al.*, *Genome-Scale Identification of Nucleosome Positions in S. cerevisiae*. Science **309**, 626 (2005).

[122] Yarrington *et al.*, *Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo*. Proc. Natl. Acad. Sci. **115**, 9351 (2018).

[123] Störtz *et al.*, *piCRISPR: Physically informed deep learning models for CRISPR/Cas9 off-target cleavage prediction*. Artif. Intell. Life Sci. **3**, 100075 (2023).

[124] Mak *et al.*, *Comprehensive computational analysis of epigenetic descriptors affecting CRISPR-Cas9 off-target activity*. BMC Genomics **23**, 805 (2022).

[125] Vinodkumar *et al.*, *Prediction of sgRNA Off-Target Activity in CRISPR/Cas9 Gene Editing Using Graph Convolution Network*. Entropy **23**, 608 (2021).

[126] Alharbi *et al.*, *nuMap: A web platform for accurate prediction of nucleosome positioning*. Genomics, Proteomics Bioinforma. **12**, 249 (2014).

[127] Zhang *et al.*, *LeNup: learning nucleosome positioning from DNA sequences with improved convolutional neural networks*. Bioinformatics **34**, 1705 (2018).

[128] Xi *et al.*, *Predicting nucleosome positioning using a duration Hidden Markov Model*. BMC Bioinformatics **11**, 346 (2010).

[129] Kato *et al.*, *Chemical map-based prediction of nucleosome positioning using the Bioconductor package nuCpos*. BMC Bioinformatics **22**, 322 (2021).

[130] Van Der Heijden *et al.*, *Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy*. Proc. Natl. Acad. Sci. U. S. A. **109** (2012).

[131] Zenil *et al.*, *Training-free measures based on algorithmic probability identify high nucleosome occupancy in DNA sequences*. Nucleic Acids Res. **47**, e129 (2019).

[132] Tennakoon *et al.*, *BatMis: a fast algorithm for k-mismatch mapping*. Bioinformatics **28**, 2122 (2012).

[133] Konstantakos *et al.*, *CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning*. Nucleic Acids Res. **50**, 3616 (2022).

[134] Lundberg *et al.*, "A Unified Approach to Interpreting Model Predictions". In: *Adv. Neural Inf. Process. Syst. 30*. 2017, 4765.

[135] Verosloff *et al.*, *RNA Sequence and Structure Determinants of Pol III Transcriptional Termination in Human Cells*. J. Mol. Biol. **433**, 166978 (2021).

[136] Doench *et al.*, *Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation*. Nat. Biotechnol. **32**, 1262 (2014).

[137] Bravo *et al.*, *Structural basis for mismatch surveillance by CRISPR–Cas9*. Nature **603**, 343 (2022).

[138] Dhanjal *et al.*, *Evaluation of off-targets predicted by sgRNA design tools*. Genomics **112**, 3609 (2020).

[139] Xu *et al.*, *Sequence determinants of improved CRISPR sgRNA design*. Genome Res. **25**, 1147 (2015).

[140] Boyle *et al.*, *Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement*. Sci. Adv. **7**, 5496 (2021).

[141] Vinogradov, *DNA helix: the importance of being GC-rich*. Nucleic Acids Res. **31**, 1838 (2003).

[142] Cofsky *et al.*, *CRISPR–Cas9 bends and twists DNA to read its sequence*. Nat. Struct. Mol. Biol. **29**, 395 (2022).

[143] Xiang *et al.*, *Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning*. Nat. Commun. **12**, 3238 (2021).

[144] Ester *et al.*, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Proc. 2nd Int. Conf. Knowl. Discov. Data Min. (1996).

[145] Katti *et al.*, *CRISPR in cancer biology and therapy*. Nat. Rev. Cancer **22**, 259 (2022).

[146] Bae *et al.*, *Microhomology-based choice of Cas9 nuclease target sites*. Nat. Methods **11**, 705 (2014).

[147] Overbeek *et al.*, *DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks*. Mol. Cell **63**, 633 (2016).

[148] Chakrabarti *et al.*, *Target-Specific Precision of CRISPR-Mediated Genome Editing*. Mol. Cell **73**, 699 (2019).

[149] Macke *et al.*, *Modeling Unusual Nucleic Acid Structures*. ACS Symp. Ser. **682**, 379 (1998).

[150] Minary *et al.*, *Training-free atomistic prediction of nucleosome occupancy*. Proc. Natl. Acad. Sci. **111**, 6293 (2014).

[151] Rohs *et al.*, *Unraveling Proteins: A Molecular Mechanics Study*. Biophys. J. **76**, 2760 (1999).

[152] McVey *et al.*, *MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings*. Trends Genet. **24**, 529 (2008).

[153] Thyme *et al.*, *Internal guide RNA interactions interfere with Cas9-mediated cleavage*. Nat. Commun. **7**, 11750 (2016).

[154] Jiang *et al.*, *CRISPR-Cas9 Structures and Mechanisms*. Annu. Rev. Biophys. **46**, 505 (2017).

Appendix A

Work on other publications

A.1 Analysis of epigenetic descriptors

Under the title "Comprehensive computational analysis of epigenetic descriptors affecting CRISPR-Cas9 off-target activity", I have contributed to a publication in which we utilised the crisprSQL database to analyse the effect of 6 experimental epigenetic and 13 nucleosome organization-related computed features on off-target activity [124]. We augmented the off-target dataset with experimental MNase-seq data and computed nucleosome organisation-related scores and made this available to the research community at https://crisprsql.com/downloads/260520_putative_nucleosomal.parquet.gz.

We were able to show that computed features (based on BDM or NuPoP models) characterising nucleosome organisation have a considerably higher positive correlation with experimentally verified off-target cleavage activity when compared to the six experimental epigenetic features (CTCF, DNase I, DRIP, H3K4me3, MNase and RRBS), four of which are frequently used in state-of-the-art deep learning-based off-target activity prediction models. After training purely sequence-dependent cleavage prediction models, we used SHAP summary plots to show that unlike the six experimental epigenetic scores, GC147, Nucleotide BDM and NuPoP(Affinity) are the top three computed nucleosome organization-related features contributing most to the model output. The publication is reprinted here.

Mak *et al. BMC Genomics* (2022) 23:805
https://doi.org/10.1186/s12864-022-09012-7

BMC Genomics

RESEARCH

Open Access

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

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

Abstract

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

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

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

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

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

Background

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

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

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

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

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

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

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

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

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

Results

Spearman and Pearson correlation analysis

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

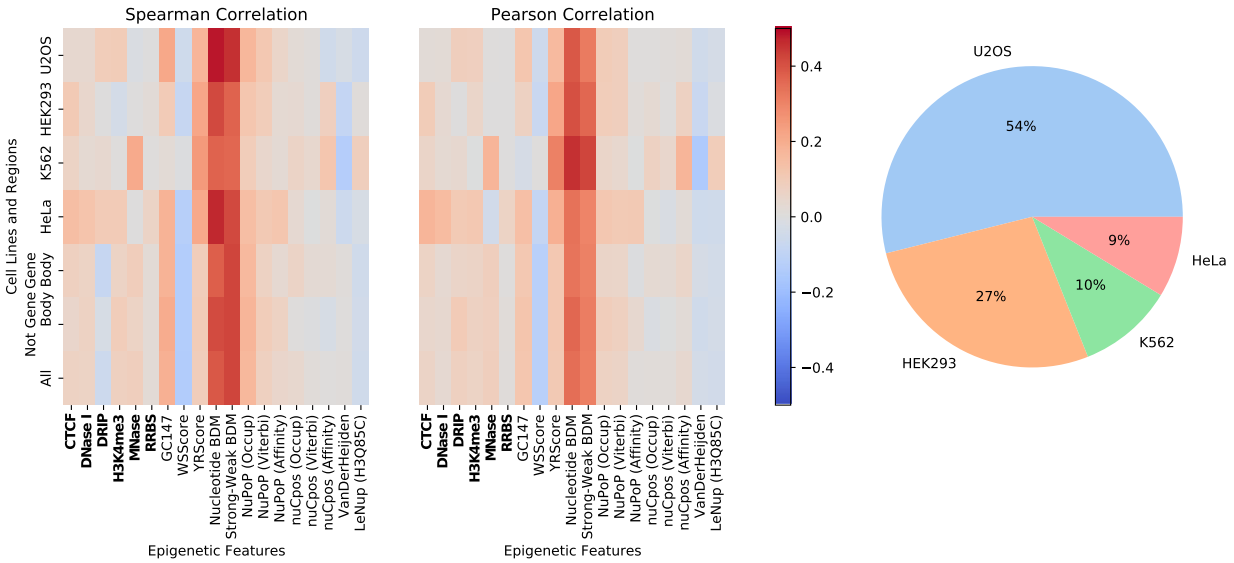


Fig. 1 Heatmaps showing Spearman (left) and Pearson (middle) correlations between 19 epigenetic features and SpCas9 off-target cleavage activities. Red and blue colors represent positive and negative correlations, respectively. The 19 epigenetic features consists of six experimental epigenetic features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

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

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

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

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

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

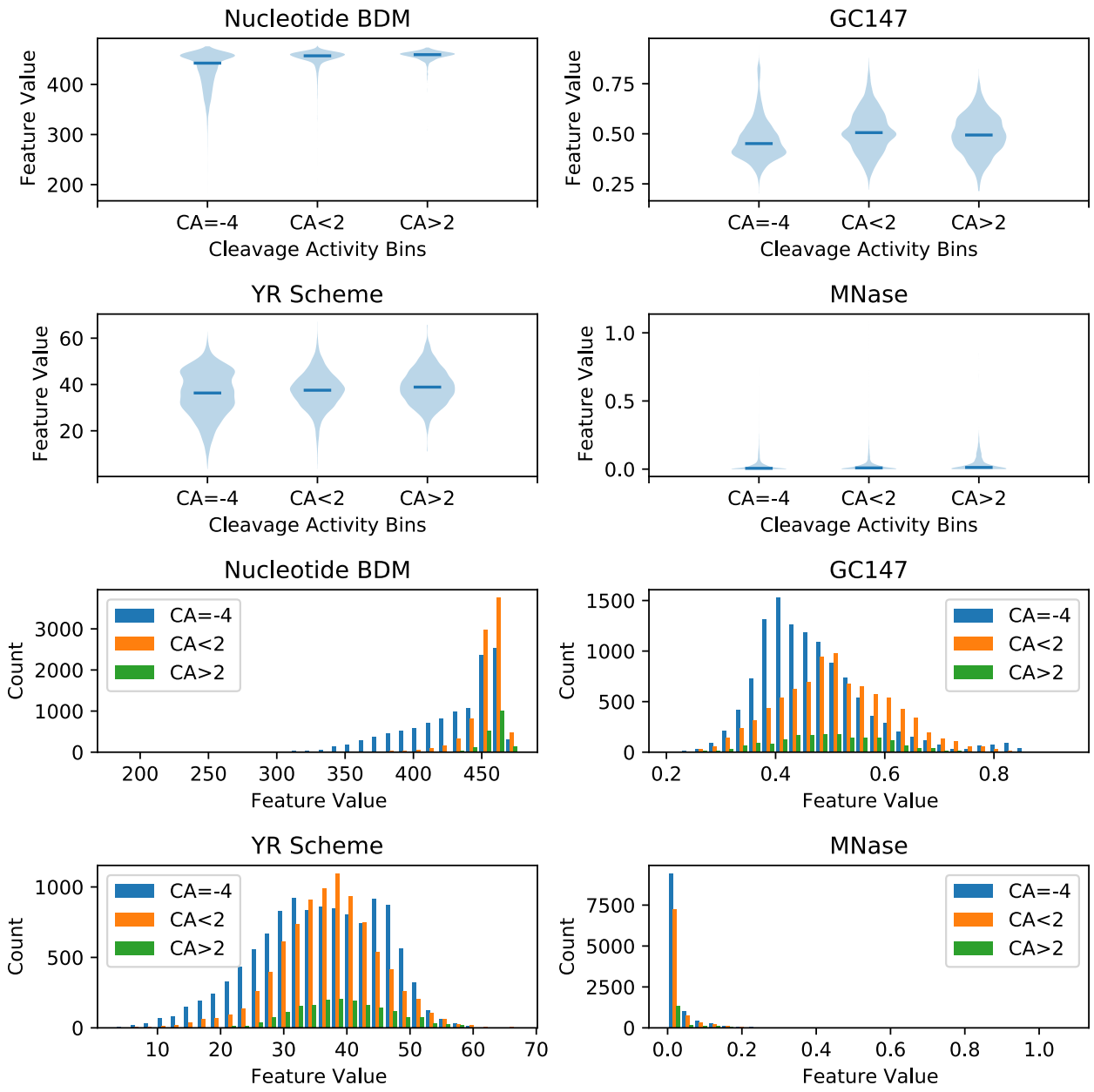


Fig. 2 Violin (top) and distribution (bottom) plots for the epigenetic features with high Pearson correlation, namely Nucleotide BDM, GC147, YR Scheme and MNase. Cleavage activities (CA) are separated into three bins representing low (CA $= -4$, colored blue), medium (CA ≤ 2, colored orange) and high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

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

Machine/Deep learning-based SHAP analysis

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

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

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

Discussion

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

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

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

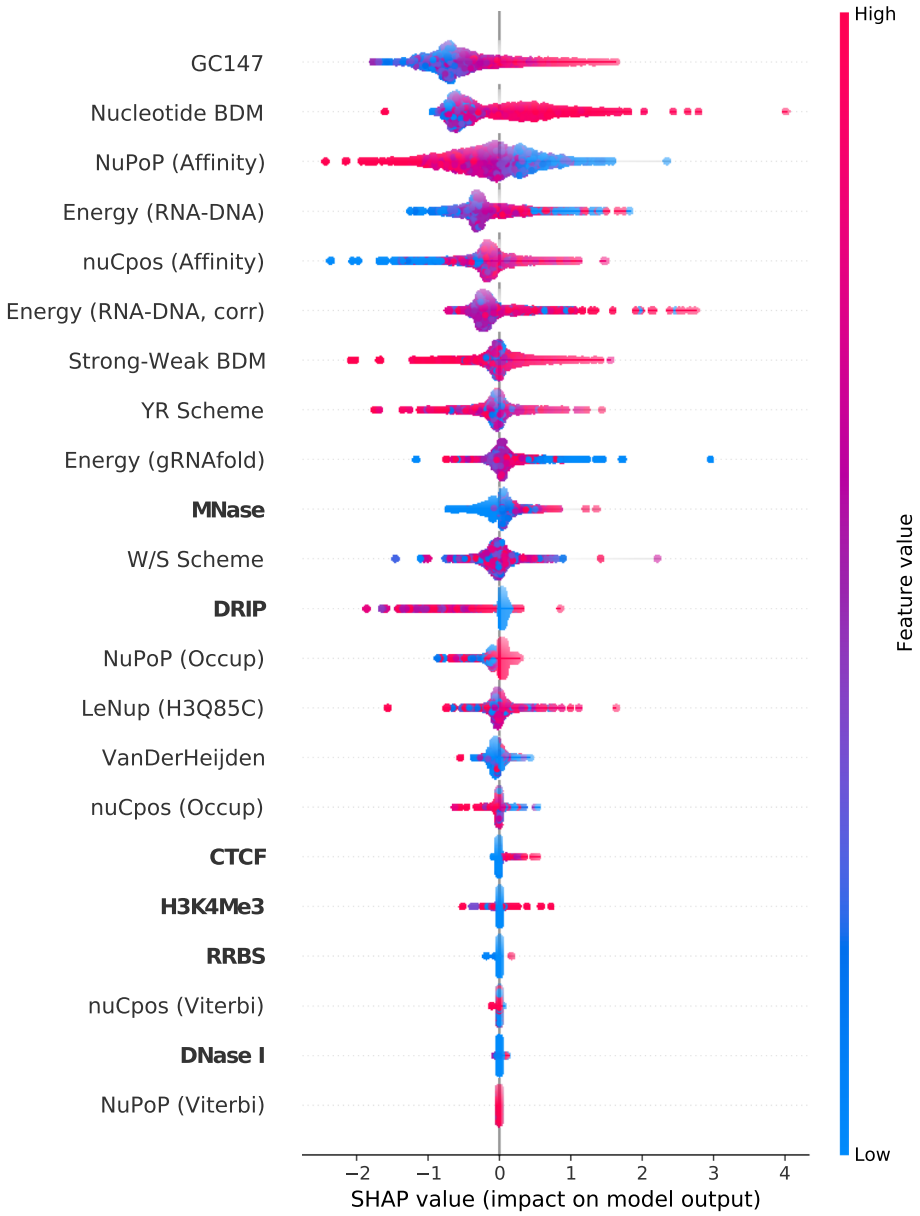


Fig. 3 SHAP summary plot for the trained extreme gradient boosted (XGBoost) tree model. The model's input consists of three CRISPRspec-derived energy terms, six experimental epigenetic scores (bolded), and 13 computed nucleosome organization-related scores. The three CRISPRspec-derived energy terms are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

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

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

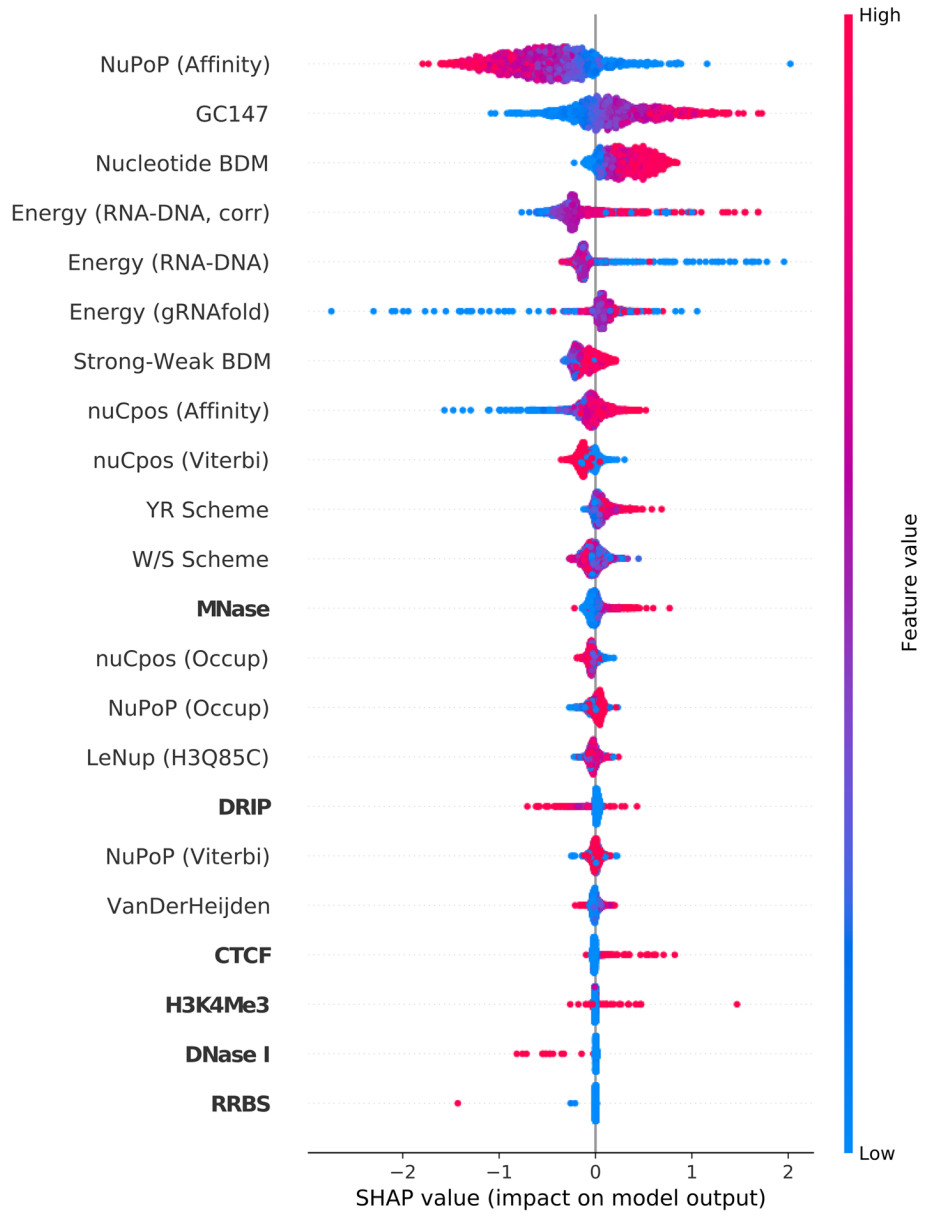


Fig. 4 SHAP summary plot for the trained convolutional neural network (CNN) model (see Supplementary Fig. 1 for architecture details). The model's input consists of three CRISPRspec-derived energy terms, six experimental epigenetic scores (bolded), and 13 computed nucleosome organization-related scores.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

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

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

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

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

Conclusions

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

Methods

crisprSQL

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

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

Experimental nucleosome occupancy data

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

Adding epigenetic scores

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

Adding nucleosome organization-related scores

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

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

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

GC content

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

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

W/S and YR schemes

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

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

Van Der Heijden algorithm

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

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

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

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

Block decomposition method-based measures

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

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

NuPoP

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

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

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

nuCpos

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

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

LeNup

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

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

Correlation and distribution analysis

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

CRISPRspec

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

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

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

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

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

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

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

Model and SHAP

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

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

Abbreviations

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

Supplementary Information

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

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

Acknowledgements

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

Authors' contributions

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

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

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

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

Declarations

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

Consent for publication
Not applicable.

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

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

References

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

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

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Appendix B

Physically informed modelling: additional data

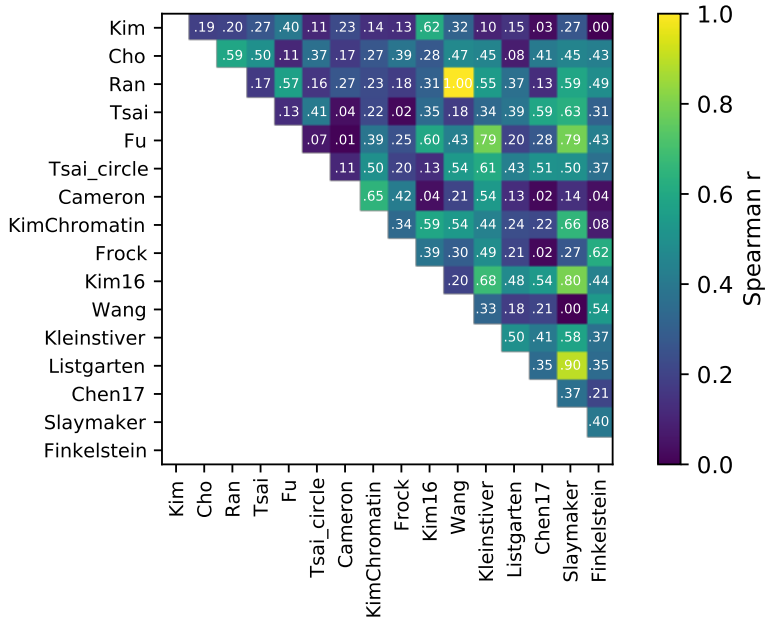


Figure B.1: Spearman r data underlying Figure 3.7. Data has been gained using the CNN_S4E0 model.

Figure B.3 shows that the RNN_S4E3 model is heavily influenced by the SW-BDM feature which is calculated from the 147 base pair region around each nucleotide of the (off-)target DNA. This stands in contrast to the implicit assumption in most cleavage prediction models that sequence context only has a minor effect on cleavage. Our results emphasise the value

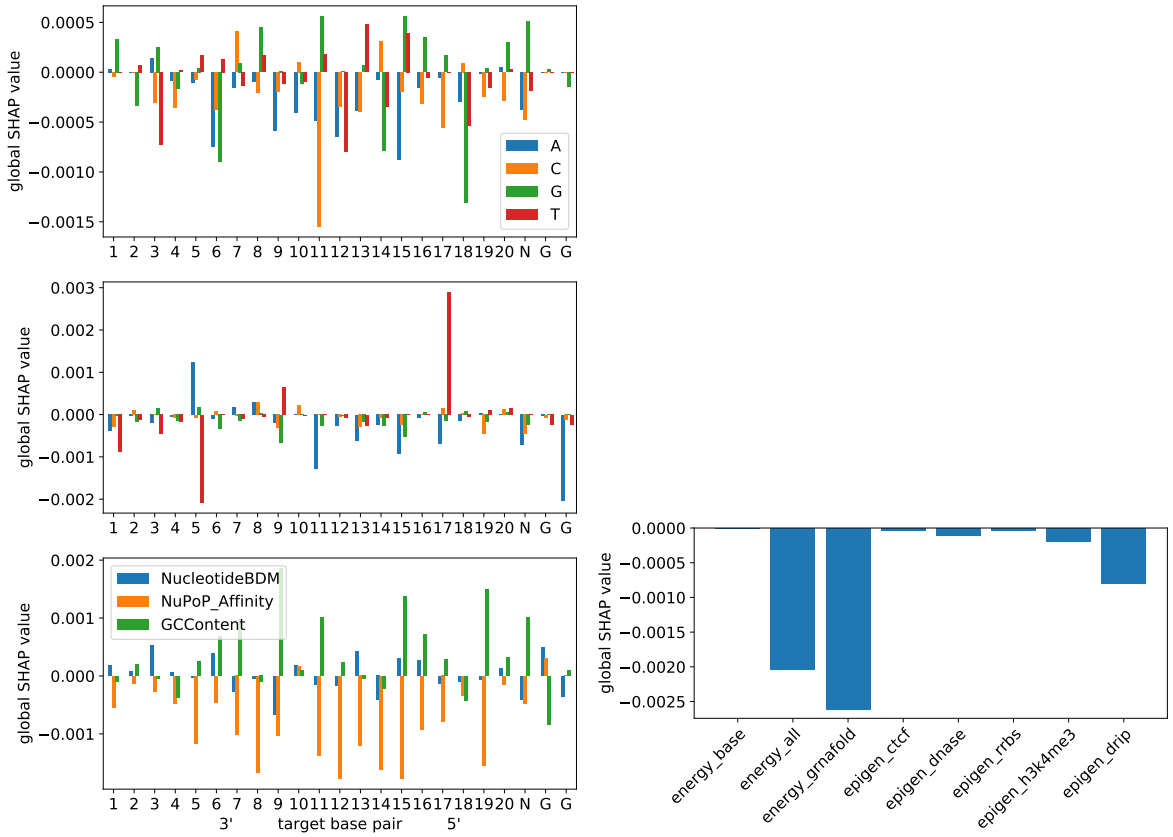


Figure B.2: SHAP values for the CNN_S5E2 model (Figure 3.12).

of the said 147bp sequence context for prediction, since the feature has been shown to carry valuable information relevant to primary chromatin structure [131].

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

Sequence features appear to have little importance in the intermediate region from positions 8 to 14, supporting the notion from [154] that PAM-distal and PAM-proximal mismatches reduce cleavage activity due to cleavage inactivation and abrogation of DNA binding, respectively. We attribute the fact that this model gives little importance to energy fea-

tures to the nature of recurrent neural networks penalising constant features across the recurrence dimension.

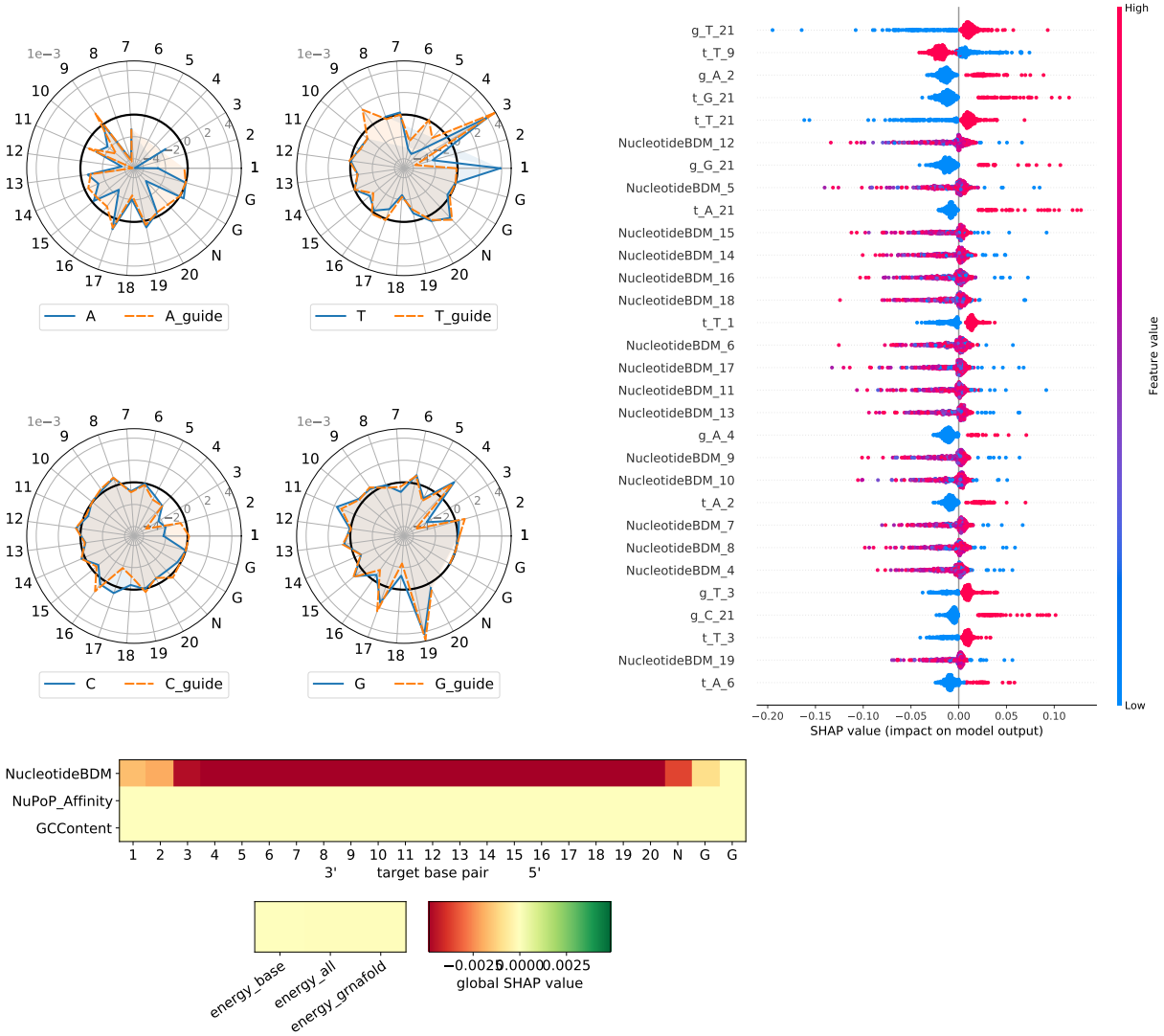


Figure B.3: Base-pair resolved global SHAP values for the RNN_S4E3 classification model. Training and test set are identical to Figure 3.12. Due to the E3 encoding which combines target and guide sequences through a bitwise OR encoding (see Figure 3.2), target (solid lines in radial plot) and guide (dashed lines in radial plot) feature importances align for most nucleotides. The heatmap shows that the StrongWeakBDM channel contributes the most to the model prediction, with PAM-proximal (distal) nucleotides showing positive (negative) correlation with cleavage activity. PAM-distal and PAM-proximal sequence features inform the model prediction more than intermediate sequence features, matching what was observed in [108]. Energy features show an order of magnitude less influence on model prediction than both sequence and nucleosomal features. A bar chart representation of this can be found in Figure B.4.

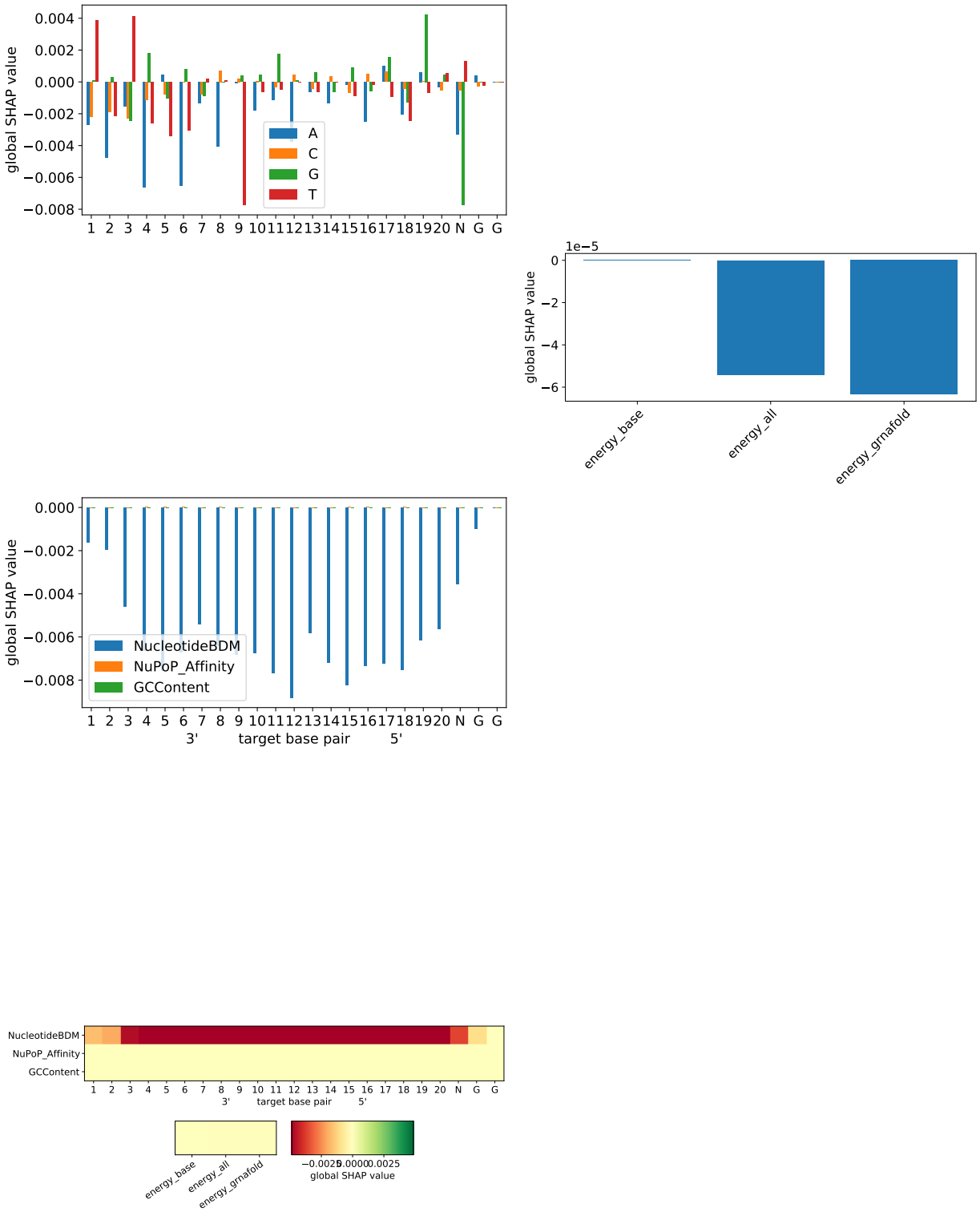


Figure B.4: SHAP values for the RNN_S4E3 classification model (Figure B.3).

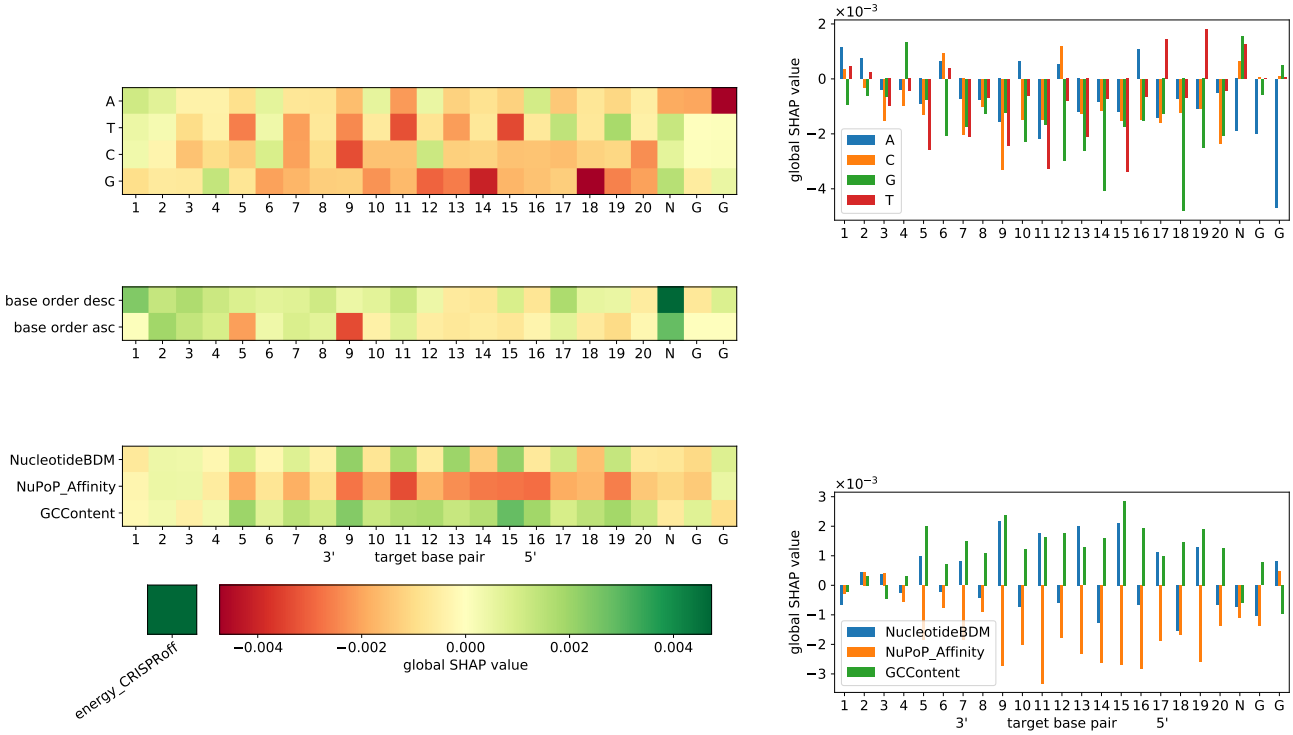


Figure B.5: Base-pair resolved global SHAP values for the 6×23 CNN classification model. Training and test set are identical to Figure 3.19. The heatmap shows that after sequence features, the NuPoP Affinity channel contributes the most to the model prediction, with PAM-distal nucleotides less important than PAM-proximal ones.

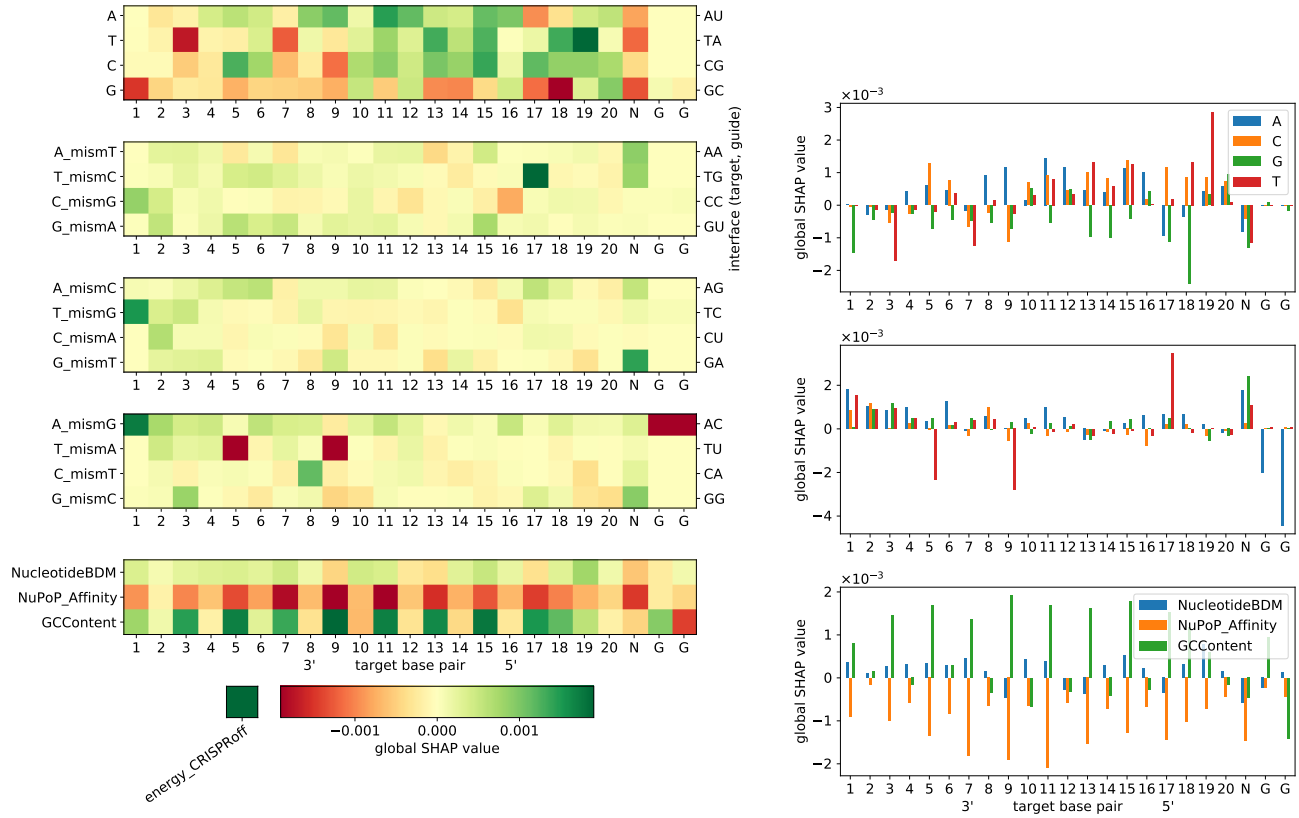


Figure B.6: Base-pair resolved global SHAP values for the 16×23 CNN classification model. Training and test set are identical to Figure 3.19. The heatmap shows that after sequence features, the NuPoP Affinity channel contributes the most to the model prediction, with PAM-distal nucleotides less important than PAM-proximal ones.

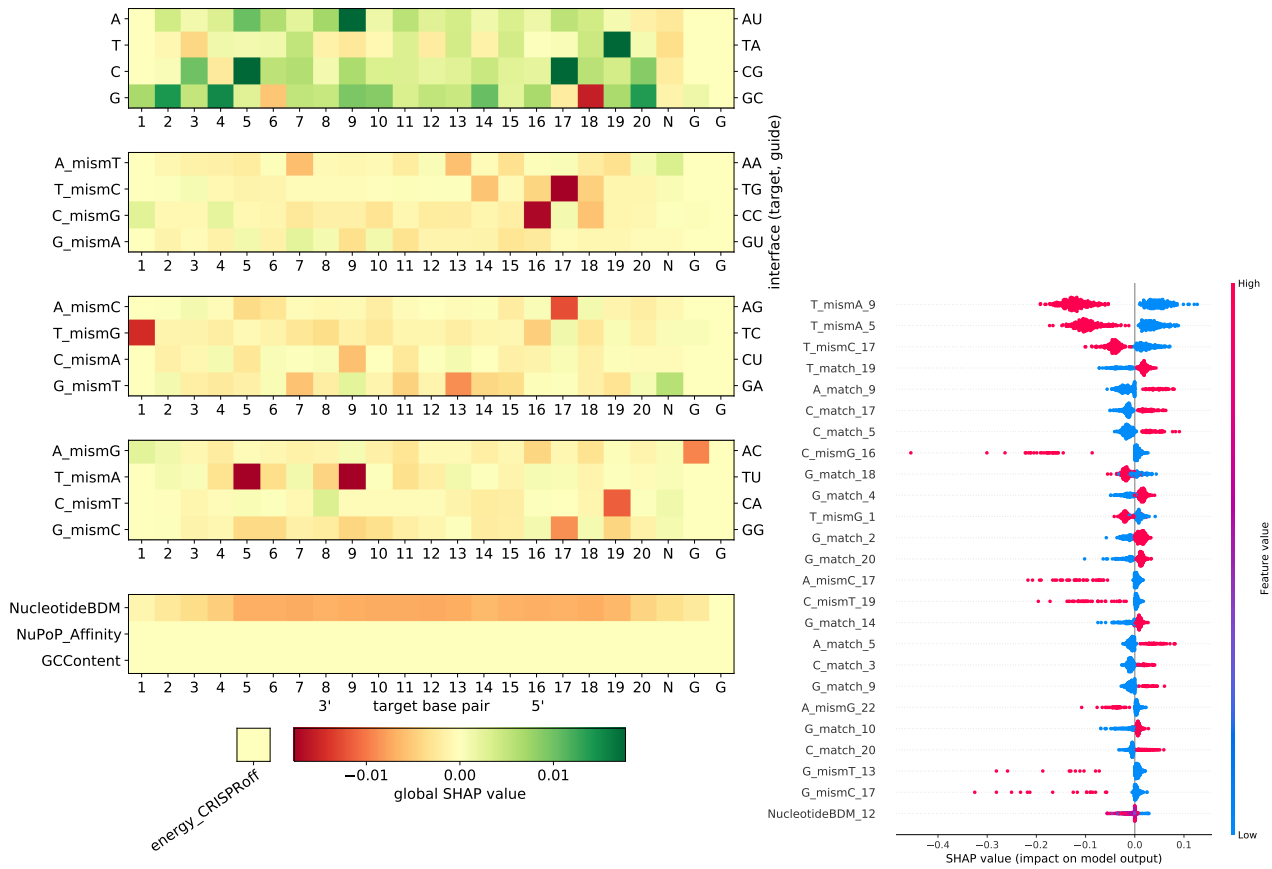


Figure B.7: Base-pair resolved global SHAP values for the 16×23 RNN classification model. Training and test set are identical to Figure 3.19. The heatmap shows that after sequence features, the Nucleotide BDM channel contributes the most to the model prediction, with PAM-distal nucleotides less important than PAM-proximal ones.

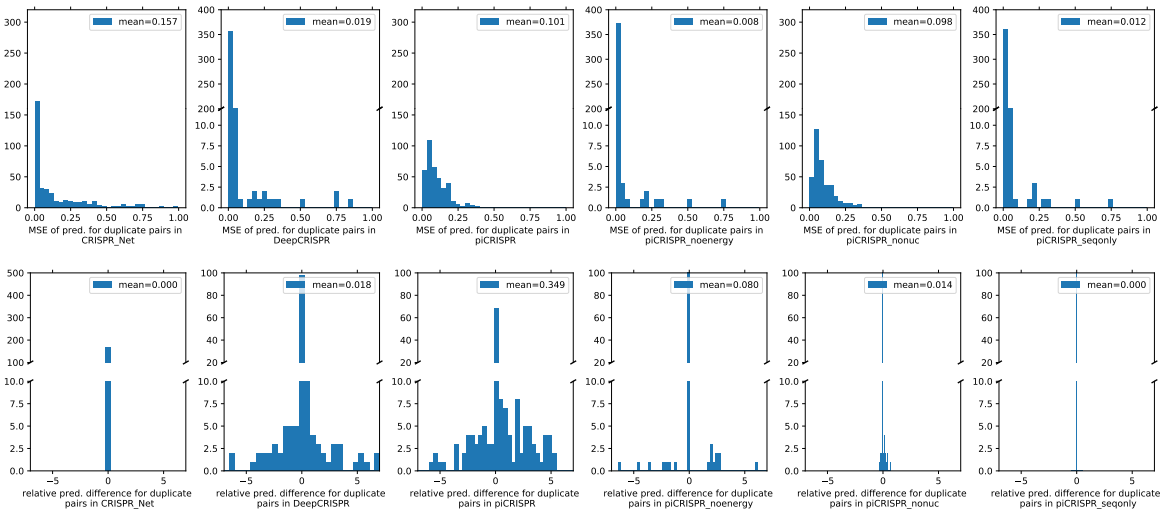


Figure B.8: Distributions underlying Table 3.3.

Appendix C

Repair outcome modelling: additional data

MOSAICS generates output of the following form:

```
================================================================================
Initial energies in kcal/mol
--------------------------------------------------------------------------------
Initial Bond energy 103.3709598637767
Initial Bend energy 1810.490018698361
Initial Torsion energy 7206.836981274822
Initial Onefour energy 751.8614890832863
Initial Inter energy 2794.404056449132
Initial Constraint energy 0
Initial Total energy 12666.96350536938
--------------------------------------------------------------------------------
Initial energies
================================================================================


================================================================================
Conjugate gradient minimization
Current Desired Conditions >> 2.0 |A-B| <= ftol (|A|+|B|+EPS) <<
{ ftol: 0.01, EPS: 1e-10 }
--------------------------------------------------------------------------------
energy 4087.68719594819 kcal/mol  ||  A = 6.514146933314636
                                      B = 6.514146933314636

--------------------------------------------------------------------------------
Desired condition is satistied
Conjugate gradient minimization terminated
```