TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Offerhaus, H. S. (2026). *From R-loops to Genomes: Connecting Molecular CRISPR-Cas9 Kinetics to Genome Editing*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b43f7e30-930b-4a7a-8a74-c1f75bff3386

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

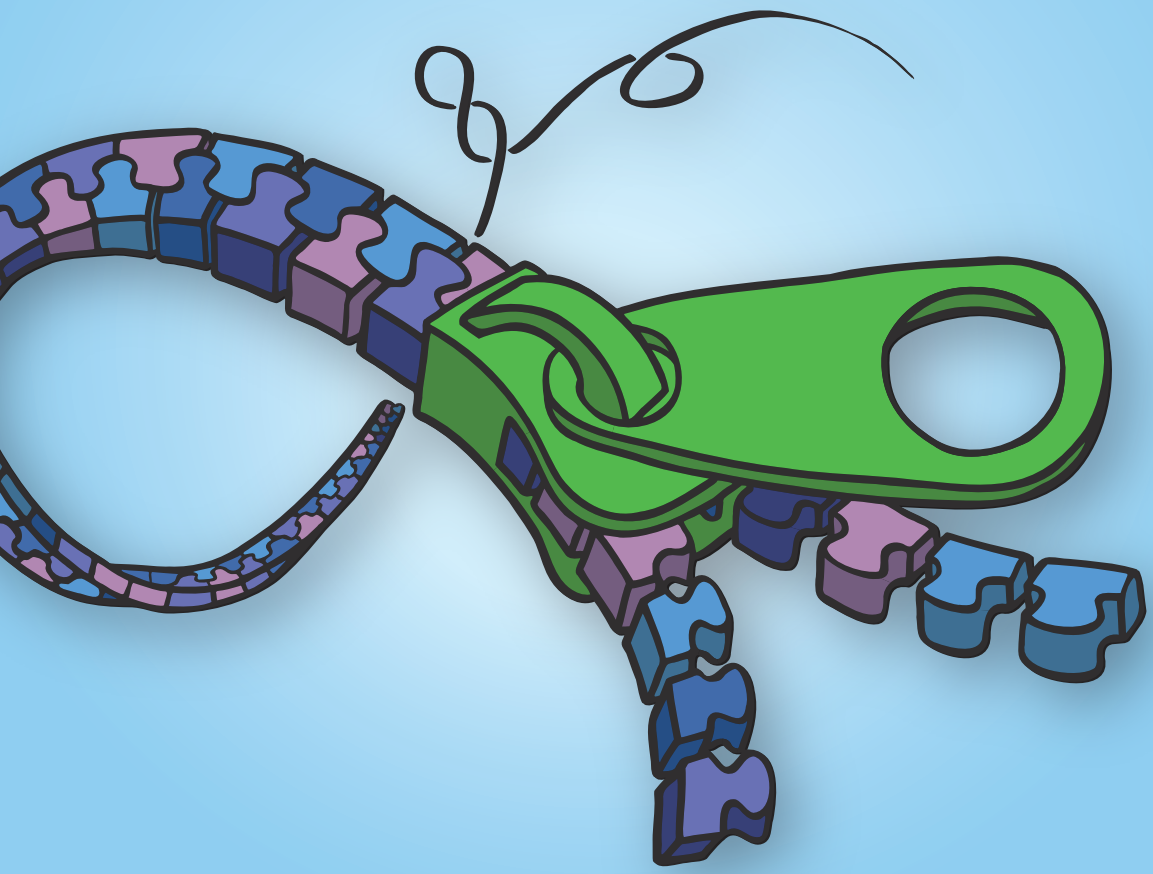
Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

From R-loops To Genomes

Connecting Molecular CRISPR-Cas9 Kinetics To Genome Editing



Hidde Offerhaus

From R-loops to Genomes

Connecting Molecular CRISPR-Cas9 Kinetics to Genome Editing

From R-loops to Genomes

Connecting Molecular CRISPR-Cas9 Kinetics to Genome Editing

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Tuesday 7 July 2026 at 10:00

by

Hidde Sieger OFFERHAUS

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Dr. S. M. Depken,	Delft University of Technology, *promotor*
Prof. dr. C. Joo,	Delft University of Technology, *promotor*
Independent members:	
Prof. dr. M. E. Aubin-Tam,	Delft University of Technology,
Prof. dr. R. P. Seidel,	Leipzig University, Germany,
Dr. J. W. Zwanikken,	Delft University of Technology,
Dr. D. Dulin,	Vrije Universiteit Amsterdam,
Prof. dr. C. Dekker,	Delft University of Technology, *reserve member*
Other members:	
Dr. S. K. Jones,	Vilnius University, Lithuania

Dr. S. K. Jones of Vilnius University, Lithuania has contributed greatly to the preparation of this dissertation.

Keywords: CRISPR-Cas9, genome engineering, molecular biophysics, enzyme kinetics, machine learning, DNA topology

Printed by: Proefschriftspecialist

Cover by: Hidde Offerhaus

Copyright © 2026 by H.S. Offerhaus

An electronic copy of this dissertation is available at https://repository.tudelft.nl/.

Contents

Summary ix

Samenvatting xi

Notation xiii

1 CRISPR: from antiviral defense to optimized gene editing 1
- 1.1 Introduction . . . 2
- 1.2 CRISPR immune systems . . . 2
  - 1.2.1 Adaptation and interference . . . 2
  - 1.2.2 The protospacer-adjacent motif (PAM) . . . 3
  - 1.2.3 *Streptococcus pyogenes* Cas9 (SpCas9) . . . 4
- 1.3 CRISPR gene editing techniques . . . 5
  - 1.3.1 Editing techniques . . . 5
  - 1.3.2 Cas variants and alternatives . . . 7
- 1.4 CRISPR applications . . . 7
  - 1.4.1 Industrial bioengineering . . . 8
  - 1.4.2 Plant biotechnology . . . 8
  - 1.4.3 Medicine and healthcare . . . 8
  - 1.4.4 Germline editing . . . 9
- 1.5 Gene editing challenges . . . 10
  - 1.5.1 Off-target activity and its consequences . . . 10
  - 1.5.2 Experimental variables . . . 10
  - 1.5.3 Genome accessibility and supercoiling . . . 12
- 1.6 gRNA design: measuring and predicting activity . . . 13
  - 1.6.1 Measuring activity *in vitro* and *in vivo* . . . 13
  - 1.6.2 Predicting activity *in silico* . . . 14
  - 1.6.3 Biophysical modeling . . . 15
- 1.7 This thesis . . . 16

References . . . 17

2 A general kinetic framework unifies Cas9 applications *in vitro* and *in vivo* 29
- 2.1 Introduction . . . 30
- 2.2 Simple kinetic model of target cleavage . . . 30
- 2.3 Rate of PAM-binding . . . 31
- 2.4 Enzymatic limits *in vivo* and *in vitro* . . . 32
- 2.5 Conclusion . . . 35
- 2.6 Supplementary Notes . . . 35
  - 2.S1 Effective cleavage rate . . . 35
  - 2.S2 RNP equilibrium concentration . . . 36

2.S3 RNP equilibration rate . . . 37
2.S4 Equilibrium concentrations for abundant and scarce RNP and PAM 39
References . . . 39

3 CRISPRzip explains and predicts sequence- and context-dependent Cas9 activity 41
3.1 Introduction . . . 42
3.2 Results . . . 43
3.2.1 Model definition . . . 43
3.2.2 Obtaining training data with high-throughput kinetic assays . . . 44
3.2.3 Establishing a single model for variable-gRNA activity prediction . 45
3.2.4 Free-energy landscapes explain variation in cleavage and binding dynamics . . . 48
3.2.5 CRISPRzip adapts to low Cas9 concentrations . . . 50
3.2.6 CRISPRzip adapts to DNA superhelicity . . . 51
3.3 Discussion . . . 53
3.4 Methods . . . 55
3.4.1 Kinetic model . . . 55
3.4.2 R-loop cost . . . 56
3.4.3 Library preparation, cleavage and binding assays . . . 56
3.4.4 Determination of binding and cleavage kinetics . . . 57
3.4.5 CRISPRzip training . . . 58
3.4.6 Binding kinetics analysis . . . 59
3.4.7 Simulation of single-molecule FRET experiments . . . 59
3.4.8 PAM-binding experiments . . . 59
3.4.9 DNA superhelicity effects on Cas9 dynamics . . . 60
3.5 Author contributions . . . 62
3.6 Acknowledgements . . . 62
3.7 Data and software availability . . . 62
3.8 Supplementary Figures . . . 63
3.9 Supplementary Data . . . 68
References . . . 72

4 Supercoiling twists Cas9 off-target discrimination when binding, nicking and cleaving 79
4.1 Introduction . . . 80
4.2 Results . . . 81
4.2.1 NucleaSeq captures Cas9 cleavage kinetics on supercoiled DNA 81
4.2.2 Mispairs affect Cas9 target cleavage rate and position . . . 81
4.2.3 Cleavage site-adjacent mispairs disrupt Cas9 nuclease activity . 83
4.2.4 A biophysical model captures the impact of supercoiling on Cas9 specificity . . . 87
4.3 Discussion . . . 89
4.4 Methods . . . 90
4.4.1 sgRNAs . . . 90
4.4.2 Plasmid library construction . . . 90

4.4.3 Plasmid library transformation and extraction . . . 90
4.4.4 Superhelical density determination . . . 90
4.4.5 NucleaSeq . . . 91
4.4.6 NGS data analysis: identification, normalization and rate fitting . 91
4.4.7 Single targets construction . . . 91
4.4.8 Single target nicking and cleavage assays and analysis . . . 92
4.4.9 Statistics . . . 93
4.4.10 Model methods . . . 93
4.5 Author contributions . . . 94
4.6 Supplementary Figures . . . 95
4.7 Supplementary Data . . . 97
References . . . 100

5 From molecular Cas9 kinetics to genome-wide cleavage patterns in cells 107
5.1 Introduction . . . 108
5.2 Results . . . 109
5.2.1 Genomes typically contain too few partial matches to delay target search . . . 109
5.2.2 Modeling Cas9 trapping experiments in *E. coli* . . . 112
5.2.3 Human genome targets are cleaved independently, without competition . . . 113
5.2.4 Cleavage capacity of Cas9 delivery determines editing outcomes 116
5.2.5 Cleavage is dominated by few strong off-targets . . . 118
5.3 Discussion . . . 120
5.4 Methods . . . 121
5.4.1 Sequence-averaged CRISPRzip predictions . . . 121
5.4.2 Off-target cleavage probabilities . . . 121
5.4.3 Prediction of genome cleavage . . . 122
5.5 Supplementary Notes . . . 122
5.S1 PAM arrival time with non-target traps . . . 122
5.S2 ODE system of genome cleavage . . . 123
5.6 Supplementary Figures . . . 125
References . . . 125

6 Towards interpretable and adaptable CRISPR activity prediction 131
6.1 Conclusions . . . 132
6.2 Outlook . . . 132
References . . . 133

Acknowledgements 137

Curriculum Vitæ 143

List of Publications 145

Summary

Editing genes has become much easier since the discovery of CRISPR-Cas proteins such as Cas9. Cas9 can be directed to cut almost any location on the genome by supplying it with a short guide RNA (gRNA) with a nucleotide sequence that matches the DNA target. With tools like Cas9, scientists can study disease, improve crops and develop therapies. Despite the apparent simplicity of Cas9-based gene editing, it remains challenging to predict if a chosen gRNA causes Cas9 to cut DNA at the intended location, and if it risks cutting other locations too. A common strategy for Cas9 activity prediction is to collect large datasets describing which gRNAs cause which edits in a particular cell, and to train machine learning models on the observed patterns. However, such empirical models rapidly lose predictive power when applied to other cellular contexts or other experimental conditions. An alternative approach is to isolate Cas9, gRNA, and DNA from a cell, measure their dynamics with high precision, and thus learn about the molecular interactions that govern Cas9 activity in any context. With this molecular knowledge, one can formulate rational and general predictions about activity across cells and experiments.

This thesis uses biophysical modeling to translate understanding of Cas9's molecular mechanisms to prediction of its genome editing activity in cells. **Chapter 1** introduces CRISPR, outlining its origin as an immune system in bacteria and archaea, its transformation into a DNA editing toolkit, and the wide range of genetic engineering applications it has enabled. It discusses the main challenges to successful CRISPR applications, and explains how biophysical models can support gRNA selection and other design choices. **Chapter 2** describes mathematically how Cas9 dynamics differ between cell-free experiments (*in vitro*) and living cells (*in vivo*) due to differences in Cas9 availability and the processes of target search and recognition. By connecting the two settings, the framework allows the findings of later chapters to be generalized across experimental and cellular conditions. **Chapter 3** introduces CRISPRzip, a model of Cas9 target recognition that integrates the physics of gRNA-DNA interactions, and thus explains the variation in activity across gRNA and DNA sequences. Moreover, CRISPRzip incorporates the effects of Cas9 concentration and DNA twisting (supercoiling) on how Cas9 binds and cuts DNA. Because CRISPRzip adapts to gRNA sequence and to environment conditions like concentration and supercoiling, it provides a flexible basis for gRNA design even when the application setting deviates from the experiments used for training. **Chapter 4** presents an experimental approach to quantify the effects of DNA supercoiling across thousands of DNA sequences. While these results can largely be explained by CRISPRzip, they also reveal a complex interplay between supercoiling and sequence that calls for further study. **Chapter 5** extends the molecular description of Cas9 to the scale of the full genome. Using analytical models of target search and numerical predictions from CRISPRzip, it shows that prediction of on-target and major off-target activity does not require explicit consideration of the surrounding genome sequence under typical conditions. Also, it formulates a theory that connects different Cas9 delivery

strategies and shows how to balance precision and efficiency in gene editing applications. **Chapter 6** concludes the thesis by summarizing the main results and presenting an outlook towards biophysics-supported CRISPR activity prediction in cells.

Samenvatting

Het bewerken van genen is veel eenvoudiger geworden sinds de ontdekking van CRISPR-Cas-eiwitten zoals Cas9. Cas9 kan vrijwel elke gewenste locatie in het genoom knippen wanneer het wordt voorzien van een korte “gids-RNA” (guide RNA, gRNA) met een nucleotidesequentie die overeenkomt met het DNA-doelwit. Met technieken zoals Cas9 kunnen wetenschappers ziekten onderzoeken, gewassen verbeteren en therapieën ontwikkelen. Ondanks de schijnbare eenvoud van genbewerking met Cas9, blijft het lastig te voorspellen of een bepaalde gRNA leidt tot een succesvolle knip op de beoogde locatie, en of het leidt tot ongewenst knippen van DNA elders. Een veelgebruikte strategie om de activiteit van Cas9 te voorspellen is het systematisch meten van de DNA-bewerkingen die verschillende gRNA’s veroorzaken in een bepaalde cel, en het vervolgens trainen van machine-learning-modellen op de waargenomen patronen. Zulke empirische modellen verliezen echter hun voorspellend vermogen wanneer ze worden toegepast in een andere cellulaire of experimentele context. Een alternatieve benadering is om Cas9, gRNA en DNA uit een cel te isoleren, hun dynamiek met hoge precisie te meten en zo inzicht te krijgen in de moleculaire interacties die de universele basis vormen van Cas9-activiteit. Met die moleculaire kennis kan men rationele en algemene voorspellingen formuleren over Cas9-activiteit in diverse cellen en experimenten.

In dit proefschrift gebruiken we biofysische modellen om de moleculaire mechanismen van Cas9 te vertalen naar voorspellingen over genbewerking in cellen. **Hoofdstuk 1** biedt een inleiding in CRISPR. Het beschrijft de oorsprong van CRISPR als immuunsysteem in bacteriën en archaea, de transformatie tot een genbewerkingstechniek, en het brede scala aan toepassingen dat hierdoor mogelijk is geworden. Het hoofdstuk behandelt de belangrijkste uitdagingen voor succesvolle CRISPR-toepassingen en legt uit hoe biofysische modellen ondersteuning kunnen bieden bij de selectie van gRNA’s en andere ontwerpkeuzes. **Hoofdstuk 2** is een wiskundige beschrijving van de verschillen in Cas9-dynamiek tussen celvrije experimenten (*in vitro*) en experimenten in cellen (*in vivo*). Door deze twee situaties met elkaar te verbinden, kunnen we in latere hoofdstukken onze bevindingen generaliseren naar andere experimentele en cellulaire omstandigheden. **Hoofdstuk 3** introduceert CRISPRzip, een model van DNA-identificatie door Cas9, dat de variatie in Cas9-activiteit tussen gRNA- and DNA-sequenties verklaart aan de hand van de fysica van gRNA-DNA-interacties. Daarnaast integreert CRISPRzip de effecten van Cas9-concentratie en DNA-torsie op de dynamiek waarmee Cas9 DNA bindt en knipt. Omdat CRISPRzip aangepast kan worden aan gRNA-sequentie en omgevingscondities zoals concentratie en DNA-torsie, is het breed inzetbaar voor gRNA-ontwerp, zelfs wanneer de toepassingscontext afwijkt van de experimenten waarop het model getraind is. **Hoofdstuk 4** presenteert een experimentele methode om de effecten van DNA-torsie op duizenden DNA-sequenties te kwantificeren. Deze resultaten worden grotendeels verklaard door CRISPRzip, maar ze tonen ook een complexe wisselwerking tussen torsie en sequentie die om nader onderzoek vraagt. **Hoofdstuk 5** breidt de moleculaire beschrijving van Cas9 uit naar de schaal van het volledige genoom. We

combineren analytische modellen van DNA-verkenning met CRISPRzips numerieke voorspellingen, en stellen vast dat men onder typische omstandigheden Cas9's knipactiviteit kan voorspellen zonder rekening te houden met de sequentie van het omliggende genoom. Daarnaast ontwikkelen we een theorie die de verschillen overbrugt tussen Cas9-bezorgingsstrategieën, en de balans beschrijft tussen de precisie en efficiëntie van genbewerkingstoepassingen. **Hoofdstuk 6** sluit het proefschrift af met een samenvatting van de belangrijkste resultaten en een vooruitblik richting biofysische voorspellingen van CRISPR-activiteit in cellen.

Notation

Selected notation, in order of appearance.

symbol	units	name	Chapter
k_{PAM} (k_{on})	s^{-1}	PAM-binding rate *rate at which an individual PAM site gets RNP-bound*	2, 3, 5
k_{unb} (k_{off})	s^{-1}	unbinding rate *rate at which an RNP unbinds a PAM*	2, 3, 5
$k_{\text{R},i}$	s^{-1}	R-loop formation rate	2
$k_{\text{clv},i}$	s^{-1}	effective cleavage rate *from PAM-binding and R-loop dynamics*	2, 3, 5
k_{a}^0 (k_{on}^0)	$\text{nM}^{-1}\,\text{s}^{-1}$	RNP-PAM bimolecular association rate	2, 3
K_D	nM	RNP-PAM dissociation constant	2, 3
k_{bnd}	s^{-1}	RNP-binding rate *rate at which an individual RNP binds any PAM*	2, 5
C		RNP copy number	2, 5
N		PAM site copy number	2, 5
$p_{\text{clv},i}$		cleavage probability *probability to cleave site i once PAM-bound*	2, 5
k_{f}	s^{-1}	R-loop extension rate	3, 4, 5
U_b	k_BT	CRISPRzip free-energy landscape	3, 4, 5
U_b^{R}	k_BT	R-loop cost	3
$U_b^{\text{P,pt}}$	k_BT	perfect-target protein contribution	3, 4, 5
Q_b^{P}	k_BT	protein mismatch penalty	3, 4, 5
α		R-loop scaling factor	3, 5
θ	° (degrees)	R-loop step angle	3, 4
φ	° (degrees)	transition angle	3, 4
σ		superhelical density	4, 5
$k_{\text{seed},i}$	s^{-1}	seed-formation rate	5
$k_{\text{reject},i}$	s^{-1}	seed-rejection rate	5
μ_i		trapping capacity	5
$\Delta U_{\text{seed},i}$	k_BT	seed stability	5
p_j	k_BT	mismatch pattern presence	5
k_{prod}	s^{-1}	production rate	5
k_{degr}	s^{-1}	degradation rate coefficient	5
k_{PAM}^1	s^{-1}	PAM-binding rate coefficient	5
Γ		cleavage capacity	5

1

CRISPR: from antiviral defense to optimized gene editing

1.1. Introduction

Since ancient times, humans have bred animals to keep and selected plants to eat, thus shaping civilization by shaping genes. As scientists uncovered the molecular basis of genetics in the second half of the twentieth century, a new possibility emerged: to directly edit the DNA underlying the traits of interest. With the right molecular tools, one would be able to edit genes throughout all forms of life, from the bacteria that produce our yogurt to the fruits that we put in it. In medicine, these tools would enable new treatments of disease that directly target the underlying genetics, possibly even personalized to a patient's unique genetic makeup [1]. To realize these future prospects, the gene editing toolbox required a pair of molecular scissors to cut genes at the precise location where DNA should be added, removed, or edited. Different generations of enzymes fulfilled this role (restriction enzymes, zinc finger nucleases, TALENs), but each required laborious redesign to target a chosen DNA site [2, 3]. In the 2010s, a simpler and more flexible solution emerged with the discovery and adoption of CRISPR–Cas enzymes [4].

1.2. CRISPR immune systems

The discovery of CRISPR-Cas enzymes is a triumph of modern biology. As high-tech as the derived techniques may seem, we remain billions of years behind the simplest life forms on Earth—bacteria and archaea—in our use of these enzymes.

1.2.1. Adaptation and interference

Throughout their existence, bacteria and archaea have been under constant attack by viruses and other hostile mobile genetic elements. As a result of the arms race dynamic between prokaryotes and their invaders, these organisms have equipped themselves with a wide variety of defense mechanisms [5]. CRISPR is one class of these immune systems, employed by 40% of bacteria and 85% of archaea [6].

Among the diverse CRISPR systems, the universal and defining feature is their specific mechanism of "adaptation" to keep a memory of previous infections. This memory takes the shape of the "CRISPR array" that the host bacterium or archaeon stores on its genome (**Figure 1.1**). It is the remarkable nucleotide sequence of this region, described as "Clustered Regularly Interspaced Short Palindromic Repeats", that gives CRISPR its name [7]. The array is built up of units, each comprising a "spacer" corresponding to a previous invader and a "repeat" with a fixed sequence. After a host has survived a new infection, it extends its memory: the CRISPR machinery selects a fragment of the invading nucleic acid (the "protospacer"), and integrates a copy of it as a new spacer with accompanying repeat in the CRISPR array [8].

When the host is infected again by a known invader, the CRISPR system carries out its "interference" mechanism, by which it identifies hostile nucleic acid and neutralizes the attack (**Figure 1.1**). Infection triggers expression of the CRISPR array, along with *cas* genes that encode for CRISPR-associated (Cas) proteins [8]. The CRISPR transcript is processed such that each spacer-repeat unit forms a "guide RNA" (gRNA), with the repeat attaining a particular secondary structure (sometimes together with other,

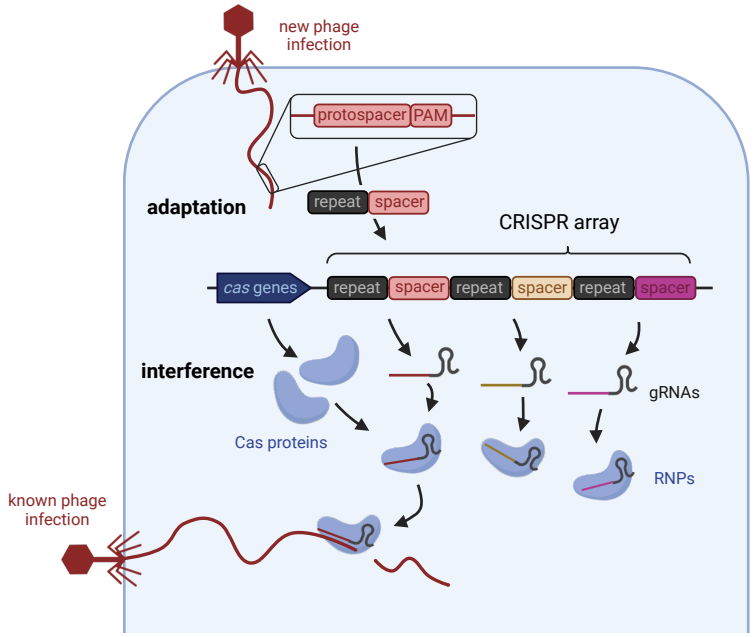


Figure 1.1: CRISPR adaptation and interference. Adaptation: after surviving a new phage infection, the CRISPR immune system selects a PAM-flanked protospacer from the invading nucleic acid, and extends the CRISPR array with a corresponding spacer and repeat. Interference: upon infection by a known phage, expression of *cas* genes and the CRISPR array creates Cas proteins and gRNAs that together form RNP effector complexes. RNPs recognize invader nucleic acid by their gRNAs' complementarity to the (PAM-flanked) protospacer, and execute their interference functionality, for example DNA cleavage.

trans-activating RNA fragments), and the spacer being left unfolded. The co-expressed Cas proteins, sometimes in complex with each other, bind the gRNAs by their folded repeat region. The formed ribonucleoprotein (RNP) complexes use the spacer region of the gRNA to recognize invaders through Watson-Crick complementarity. To do so, they bind DNA and locally unwind the helix, such that hybrid base pairs form between the gRNA and one of the DNA strands [8]. RNA-targeting Cas effectors also exist, which bind messenger RNA (mRNA) to recognize transcripts of invading DNA [3]. When a Cas effector recognizes its target, it carries out its interference functionality, which may range from cutting its DNA or RNA target, to activating other defense mechanisms, and even triggering cell death to protect the surrounding colony [6].

1.2.2. The protospacer-adjacent motif (PAM)

It is crucial that Cas effectors can distinguish the matches on foreign DNA from those on the host's own CRISPR array to prevent auto-immunity. This difference between self and non-self is marked by a protospacer-adjacent motif (PAM), a recognition sequence

(typically 2–5 bp long) directly flanking the target [9]. During adaptation, protospacers are selectively acquired from the invading genome only when they are flanked by a PAM, but in the CRISPR array itself, the neighboring PAMs are absent [8, 9]. During interference, Cas enzymes identify their intended target as such only when a PAM is present next to the DNA sequence that complements their gRNA. PAM motifs and their position with respect to the target vary among CRISPR systems [10]. A canonical example that is relevant to this thesis is the *Streptococcus pyogenes* CRISPR immune system hosting the effector protein Cas9, which features a PAM sequence 5'-NGG-3' located downstream of protospacers, with N being any nucleotide.

1.2.3. *Streptococcus pyogenes* Cas9 (SpCas9)

The protein Cas9 from the *S. pyogenes* type II-A CRISPR system (SpCas9, hereafter Cas9) is the best-known and most widely used Cas effector for gene editing. Cas9 is a DNA nuclease, which cleaves DNA upon target identification, creating a double-stranded break (DSB). Structurally, it is a relatively simple effector: it comprises a single protein, and while in nature its gRNAs are formed from multiple RNA fragments, one can synthesize a single-guide RNA (sgRNA) with a functionally similar structure [12]. The gRNA has a 5' recognition sequence that is typically 20 nucleotides long. The protein structure is characterized by two lobes [13] (**Figure 1.2A**). The recognition (REC) lobe binds the gRNA and enables DNA interrogation. The nuclease (NUC) lobe contains the two protein domains HNH and RuvC that cleave both DNA strands, as well as the PAM-interacting (PI) domain which recognizes the presence of the NGG motif.

In its search for the target, Cas9 first associates non-specifically to DNA by weakly binding the phosphate backbone with the positively charged regions of its PI domain [14] (**Figure 1.2B**). It locally diffuses along the DNA backbone until it either identifies a PAM or dissociates [15]. If the protein passes over an NGG motif, the PI domain may stably bind it by establishing specific contacts with the guanine bases [16].

Cas9 can only start to interrogate the target contents after a PAM is bound [14]. Upon PAM binding, the REC lobe closes towards the NUC lobe, which causes the DNA helix to bend and the base pairs upstream of the PAM to melt [14]. If the exposed nucleotide on the DNA target strand (TS) matches the 5' target region of the gRNA, these form a RNA:DNA hybrid base pair. This step is the first of a strand-replacement reaction, in which the hybridization of the gRNA and TS promotes further unwinding of the DNA helix, creating a so-called "R-loop" structure of a molten DNA helix with one DNA strand hybridized to RNA [17] (**Figure 1.2B**). During the process, Cas9 undergoes structural reconfigurations that are coordinated with the gRNA:TS hybridization [18, 19]. Eventually, the formation of a stable, complete R-loop allows the HNH domain to rotate towards the TS while the non-target strand (NTS) is positioned in the RuvC domain [17, 20]. The HNH and RuvC active sites cleave both DNA strands between the 3rd and 4th base pairs upstream of the PAM, leaving a blunt DSB. After cleavage, Cas9 remains stably bound to the DNA [20].

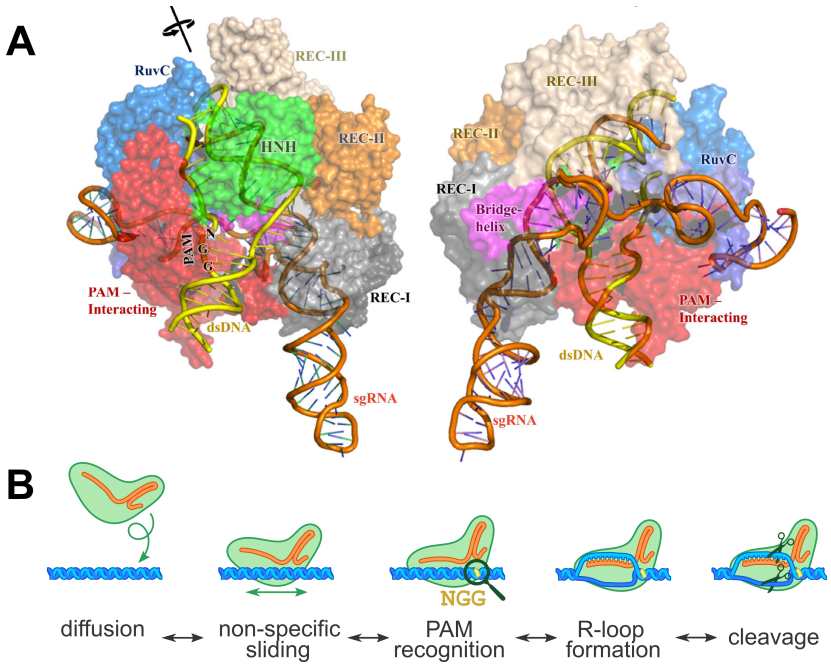


Figure 1.2: ***Streptococcus pyogenes*** **Cas9 (SpCas9). A** X-ray structure of SpCas9:gRNA in complex with DNA target, pre-cleavage (PDB 5F9R). Right structure is rotated by 180° with respect to left structure. Nucleic acid and protein domains are colored and labeled: sgRNA (orange), DNA (yellow) with PAM specified (left panel); REC-I (gray), REC-II (orange), and REC-III (beige), and Bridge-helix (magenta) of the REC lobe; RuvC (blue), HNH (green), and PAM-interacting domain (red) of the NUC lobe. Reproduced from Bhattacharya & Satpati (2023) [11], licensed under CC BY-NC-ND 4.0. **B** Schematic of Cas9 target search, PAM recognition, R-loop formation and cleavage, with nucleic acid and protein labeled as Cas9 (green), gRNA (orange), DNA TS (light blue), DNA NTS (dark blue), PAM (yellow).

1.3. CRISPR gene editing techniques

In 2012, it was demonstrated that Cas9 can be directed to target any particular DNA site with custom synthetic gRNAs [12]. With the introduction of this technique, programmable DNA cleavage no longer required custom protein engineering of restriction enzymes, zinc finger enzymes or TALENs [2], but could be achieved with simple, fast and cheap RNA synthesis [3]. Since then, the CRISPR toolkit for genome editing has expanded far beyond wild-type Cas9, now including diverse techniques and RNP effectors [21]. These tools have brought the prospect of precise DNA editing within tangible reach.

1.3.1. Editing techniques

One class of CRISPR techniques is DSB-based genome editing. In these, one uses a Cas nuclease to introduce a DSB on a genomic target, which triggers the cellular DNA repair machinery [22]. In the process of restoring the DNA helix, the repair machinery

may create desired sequence edits to the target site [23].

The first type of DSB-based edits are gene knockouts. DSBs are often repaired by non-homologous end joining (NHEJ), in which the DNA ends are fused together [22]. This error-prone process can introduce indel mutations, inserted or deleted base pairs, which may render a gene dysfunctional by disrupting its reading frame, the 3-nucleotide codon pattern [24]. Many successful gene editing applications across the life sciences are based on gene knockout (**Section 1.4**). Moreover, this technique is widely used in functional genomics, in which genotype-phenotype relations are studied based on sequencing data [25]. Here, genome-wide CRISPR knockout screens systematically disrupt the genome at different loci, allowing the identification of essential genes based on cell survival [26, 27]. In techniques like Perturb-Seq [28], a knockout screen is followed up by single-cell RNA sequencing to quantify effects on gene expression.

A second DSB-based editing method produces gene corrections or insertions. This technique relies on the homology-directed repair (HDR) pathway, which normally restores the original sequence around the DSB with the use of a DNA template containing the same DNA sequence [22, 23]. By providing the HDR machinery with a donor DNA template containing a corrected sequence or a gene knock-in, these edits are incorporated at the DSB site. While HDR-based gene editing has been demonstrated experimentally, its application is limited by relatively low success rates due to poor control over HDR repair pathways [22, 23, 29] (**Section 1.5.2**).

Another class of CRISPR techniques is DSB-free editing, utilizing Cas enzymes only to recognize and bind a target. These techniques use mutated Cas9 variants that do not perform full cleavage—catalytically dead Cas9 (dCas9) or DNA-nicking Cas9 (nCas9)—which are then fused to proteins with a particular functionality. In CRISPR interference and activation (CRISPRi/CRISPRa), dCas9 is coupled to a repressor or activator and programmed to target a particular gene of interest, modulating its expression [30, 31]. CRISPRi and CRISPRa are common techniques in functional genomics [27], but they are also applied as genome engineering tools to create phenotypic traits without permanent genetic edits [3].

The DSB-free techniques of base and prime editing do edit the genome contents [21]. In base editing, dCas9 or nCas9 is fused to a deaminase enzyme that directly converts one nucleotide into another [32]. These nucleotide conversions are resolved through endogenous DNA repair, resulting in the desired single point mutation in a gene. In prime editing, nCas9 is coupled to a reverse transcriptase and equipped with a special prime editing gRNA (pegRNA) that encodes the desired gene edits in a template at its 3' side [33]. Upon DNA nicking, one strand hybridizes with the end of the pegRNA template, after which the reverse transcriptase extends the DNA strand to match with the rest of the template. Again, DNA repair resolves the incorporation of the new DNA sequence. This system can install small insertions, deletions and point mutations. Base and prime editing have been widely adopted as an alternative to HDR-based gene editing, however they do not yet produce sufficiently reliable editing outcomes for some real-world applications, particularly for therapeutics [1, 21].

1.3.2. Cas variants and alternatives

For all of the above techniques, the protein Cas9 is a popular choice as it is well-characterized, effective across many cell types, and widely available. However, alternative nucleases may be preferred because they are smaller, are native to the studied organism, have increased editing precision, or alter the DNA repair response [21]. Also, a nuclease's specific PAM requirement limits which DNA loci can be edited. For example, to target a sequence that is not flanked by the motif NGG, a different nuclease than wild-type Cas9 is required.

Well-established wild-type CRISPR nucleases beyond Cas9 include the single-protein effector Cas12a and the protein complex Cascade [3]. Smaller Cas9 orthologues can be found in the CRISPR immune systems of *Staphylococcus aureus* (SaCas9), *Neisseria meningitides* (NmCas9) and *Campylobacter jejuni* (CjCas9), however, these enzymes have more complex PAM requirements [3]. Recent research also presented RNA-guided nucleases from transposons rather than CRISPR: the protein TnpB, a likely evolutionary ancestor of Cas nucleases, [34] and Fanzors, which are widespread in eukaryotes [35, 36]. Though gene-editing efforts based on these enzymes are still experimental, they could potentially simplify cellular delivery thanks to their small size. To target RNA, for instance for translational regulation, the nuclease Cas13 is available [37], as well as non-CRISPR RNA-guided RNA nucleases from the Argonaute family [38].

There are also a large number of engineered Cas9 variants with improved performance [39]. These are obtained either by rational protein engineering, in which the peptide sequence is adjusted, or by directed evolution, in which proteins are repeatedly randomly mutated and selected based on performance. Many variants exist with increased targeting precision, including the popular variants Cas9-HF1, [40] eSpCas9 [41] and Hypa-Cas9 [42]. Alternatively, the gRNA secondary structure may be redesigned for improved Cas9 performance [43]. Other variants are developed to relax Cas9's native PAM-requirement of NGG and extend the range of targetable sites. The variant SpRYCas9 is a near-PAMless variant which stably binds all motifs NAN and NGN, and more weakly binds motifs NCN and NTN as well [44]. Recent work has presented a different approach to flexible PAM recognition by using machine-learning based protein design to develop Cas9 variants with custom PAM motifs [45].

1.4. CRISPR applications

Within little more than a decade since its adoption as a genome engineering tool, CRISPR-based gene editing has moved from proof-of-principle experiments to widespread applications. The impact on basic research and genetics has been particularly fast and profound: countless studies have benefited from the CRISPR gene editing paradigm, in which genome manipulation and gene function analysis is simple, fast and inexpensive. Meanwhile, the techniques have enabled applications beyond fundamental research, with particularly large impact in the fields of industrial bioengineering, plant biotechnology, and medicine and healthcare. As these applications continue to advance rapidly, CRISPR technologies appear in more and more aspects of our daily lives.

1.4.1. Industrial bioengineering

Industrial bioengineering exploits the metabolism of cells to produce chemicals, materials, fuels, enzymes, and biopharmaceuticals such as vaccines. The central challenge in this field is to optimize metabolism and other cell processes for efficient, robust and scalable production. Before the introduction of CRISPR technologies, this design process was characterized by laborious low-throughput gene editing, largely based on trial-and-error [2, 46]. Now, CRISPR screens and CRISPR gene editing or gene regulation allow rapid modulation of cellular pathways and processes, resulting in a workflow of fast and iterative design-build-test-learn cycles [47, 48]. Examples of successful CRISPR applications used in industry include the production of therapeutic antibodies and hormones by mammalian cells [49], the production of biofuels by bacteria, fungi and microalgae [50], and the optimization of microbial fermentation [51].

1.4.2. Plant biotechnology

Humans have domesticated crops ever since the agricultural revolution, but in the 1980s plant scientists started to directly modify plant genomes in an effort to create genetically modified (GM) crops with higher yield, nutritional value and resistance to pests and environmental pressures. These earlier efforts included transforming plants with foreign genes, as well as pre-CRISPR targeted gene editing [2], both with low control, efficiency and throughput. Another approach was random, genome-wide mutagenesis by chemicals or radiation, followed by iterative screening for desired traits and subsequently breeding these traits into functional plants. With initially Cas9- and Cas12-based gene editing [52], and more recently, advanced CRISPR tools like base and prime editing [53], rational plant editing has become increasingly feasible through functional genomics screens and precise gene knock-outs, knock-ins, and promoter editing [54]. CRISPR GM crops include multiple varieties of high-yield rice [53], newly domesticated wild tomatoes with improved nutritional value [55], and maize variants with improved drought tolerance [53]. The CRISPR gene-editing toolkit has great potential to further stimulate future plant engineering, but commercial viability of GM crops is limited in many countries due to restrictive legislation [56] and skeptical popular opinion [57].

1.4.3. Medicine and healthcare

CRISPR tools have diverse applications across medicine and healthcare, also beyond human gene therapies. CRISPR gene editing has simplified the development of personalized disease models by introducing patient-specific gene variants into stem cells [58] and organoids [59]. Another established application is CRISPR diagnostics, making use of the ability of Cas effectors to recognize DNA and RNA sequences of pathogens and markers of cancer [60, 61]. Several CRISPR diagnostics platforms have progressed beyond proof of concept, for instance for COVID-19 testing, thanks to their minimal instrumentation requirements and rapid adaptability to new targets compared to existing diagnostic tools [62]. In the domain of public health, CRISPR-based “gene drive” efforts could potentially change population-level gene pools, for instance in malaria-transmitting mosquitoes [63]. However, ecological risks, ethical concerns, and

policy issues need to be addressed before this technology can advance from the experimental stage to real-world applications.

Despite strict safety requirements and long development timelines, the first CRISPR-based human gene therapies have recently received regulatory approval and have become available [1, 64]. The most promising clinical applications to date are treatment of genetic disorders and immune cell therapies.

Genetic disorders, particularly those with a single genetic origin, may be corrected either by *in vivo* delivery of CRISPR gene editors directly to the patient's tissue, or by *ex vivo* treatment of cells extracted from the body. In 2023, the FDA approved the first *ex vivo* therapy, which treats sickle cell disease by knocking out a repressor of hemoglobin in bone marrow stem cells [65]. In 2024, the first *in vivo* therapy advanced to late-stage clinical trials, treating the rare hereditary disorder amyloid transthyretin (ATTR) amyloidosis by knocking out the pathogenic TTR gene involved in liver metabolism [66]. A second promising *in vivo* therapy in the liver, currently in early-stage trials, targets hereditary angioedema (HAE) as it restores blood-plasma regulation through knock-out of a key pathway regulator [67]. Other recent clinical trials show potential for localized CRISPR editing in the eye to treat genetic retinal diseases [68].

Immune cell therapies involve *ex vivo* editing of bone marrow stem cells or T cells to enhance the immune response to tumors or infections [69]. In oncology, early trials have demonstrated the application of CRISPR to engineered, tumor-specific T cells to knock out immune checkpoint genes that inhibit antitumor activity [70]. Clinical trials in infectious disease show that CRISPR can knock out a T cell co-receptor required for HIV entry, generating resistance to HIV infection [71].

Beyond the above applications, more experimental CRISPR-based strategies aim to control infectious diseases by directly targeting and disabling viral genomes within infected cells [72, 73]. Looking forward, CRISPR-based therapies are being explored for more common diseases with a complex genetic component, including cardiovascular and metabolic disorders [74].

1.4.4. Germline editing

One of the most controversial potential applications of CRISPR technology is germline editing, i.e. engineering of the heritable genome in embryos or germ cells. Germline editing creates possibilities on a spectrum from preventing heritable disease to improving offspring cosmetically, intellectually or physically, with no clear boundary between therapeutics and enhancement [75]. Even in a strictly therapeutic context, there is broad scientific and ethical opposition to germline editing, as these edits can spread across many generations, with unpredictable effects [76]. The prospect of enhancement applications meets universal objection across scientists, policymakers, legal experts and the general public, driven by concerns about reinforced social inequality, discrimination and eugenic thinking. In conclusion, although proof-of-concept studies demonstrate its technological feasibility [77, 78], germline editing remains a subject of ethical discussion and public policy rather than active research.

1.5. Gene editing challenges

The successful applications listed above showcase the power of CRISPR-based gene editing, however, the toolkit comes with challenges that limit current and future applications. CRISPR applications often require careful gRNA choice and assay design to successfully achieve a particular edit in a particular cell. In the design of a CRISPR experiment, there are two primary measures of editing success to be considered: the efficiency, i.e. how effectively the intended edit is achieved, and the specificity, i.e. the risk of unintended off-target edits. While low efficiency renders an application ineffective, low specificity or high off-target activity can be toxic to the cell.

1.5.1. Off-target activity and its consequences

Early Cas9 experiments in human cells already established that some CRISPR applications cause large-scale off-target cleavage across the genome [79, 80]. Off-target cleavage typically occurs on DNA sequences that closely resemble the gRNA, specifically when the mismatches between DNA and gRNA are located distal to the PAM [81]. Cas9 also displays reduced activity on the non-canonical PAM motifs NGA and NAG [10].

From an evolutionary perspective, off-target promiscuity could be beneficial for a CRISPR immune system, as it allows recognition of viral DNA fragments that have undergone minor mutations. However, in CRISPR applications, particularly human therapies, off-target DSBs can have serious and harmful consequences [82]. First, there is the risk that a Cas nuclease knocks out tumor suppressors [83] or other essential genes [84]. These genes could also be silenced if the enzyme stably binds to them or their promoters [85, 86]. This silencing effect requires only binding and not cleavage, so it may go undetected in off-target cleavage screens, and it may be a risk even in DSB-free applications. Second, high cleavage activity beyond the intended target increases the risk of structural damage to the genome, including large insertions or deletions, chromosomal rearrangements, and even whole chromosome loss [87–89].

1.5.2. Experimental variables

In an effort to optimize the efficiency and specificity of an assay, researchers may consider several variables to CRISPR experiment design. This and later sections discuss the popular nuclease Cas9, however, the design principles apply more generally across Cas nucleases and other RNA-guided effectors.

Target site and gRNA sequence

Every application requires careful selection of a target site and corresponding gRNA, which affect editing outcomes in multiple ways. First, the sequence of the DNA target and gRNA affect the intrinsic cleavage activity of Cas9, independent of the cellular context. Nucleotide sequence is a main determinant to the molecular dynamics of R-loop formation [90]. Also, some gRNA sequences may cause unstable or deviant gRNA folding, disrupting Cas9:gRNA complex formation and lowering overall activity [91]. Next, the target site determines where on the genome Cas9 should operate. The

local DNA environment around the target site forms an extrinsic factor to Cas9 activity and editing outcomes, as further discussed in **Section 1.5.3**. Finally, gRNA sequence determines which other sites on the genome qualify as candidates for off-target activity by Cas9.

Nuclease delivery

The total cleavage activity scales with the total presence of functional Cas9:gRNA RNP in the cell or nucleus. *In vivo* delivery strategies can be classified according to their delivery format, either DNA or RNA encoding for Cas9 and gRNA, or preassembled RNPs. For either option, different delivery vehicles are available, including viruses or virus-derived vehicles, lipid nanoparticles, electroporation, microinjection, and more [92, 93]. Different delivery strategies result in different Cas9 dosage, onset timing, and activity duration [94]. Moreover, delivery strategies may trigger different cellular responses, for instance immune reactions that impede nucleic acid uptake, downregulate expression, or degrade the foreign Cas9 complexes. The delivery strategy largely determines the global activity across the genome, such that improved delivery often increases efficiency while decreasing specificity.

DNA repair

CRISPR engineering relies on DNA repair to incorporate edits, from the DSB-based methods that utilize NHEJ or HDR [23], to base and prime editing in which edits in one DNA strand cause a correction of the other strand [21]. Particularly in the context of HDR editing in eukaryotes, researchers have explored how to exert control over the DSB response. Repair pathways are tightly regulated as a function of cell cycle [22]: NHEJ occurs in all cell phases, whereas HDR generally activates in the S and G_2 phases, during which the genome copies can serve as DNA templates [22, 95]. If required, experiments can enhance HDR by optimizing donor DNA delivery [96] or suppressing the competing NHEJ pathway [97, 98]. Other factors that influence the outcomes of DNA repair are the stability with which the RNP complex binds the DNA site post-cleavage; the type of DSB being either blunt (as generated by Cas9) or overhanging (as generated by Cas12a); DNA sequence and structure around the DSB; and cellular response to delivery [22].

Nuclease choice

As discussed in **Section 1.3**, there are enzymes available beyond wild-type Cas9 which can improve editing success. High-fidelity variants are designed to increase specificity, yet often at the cost of efficiency [99]. Other wild-type Cas effectors or RNA-guided nucleases may be utilized to improve delivery (SaCas9), manipulate DNA repair (Cas12a) or reduce the cellular response to delivery (native enzymes). Finally, PAM-relaxed variants can be employed to access target sites that are not flanked by the NGG motif.

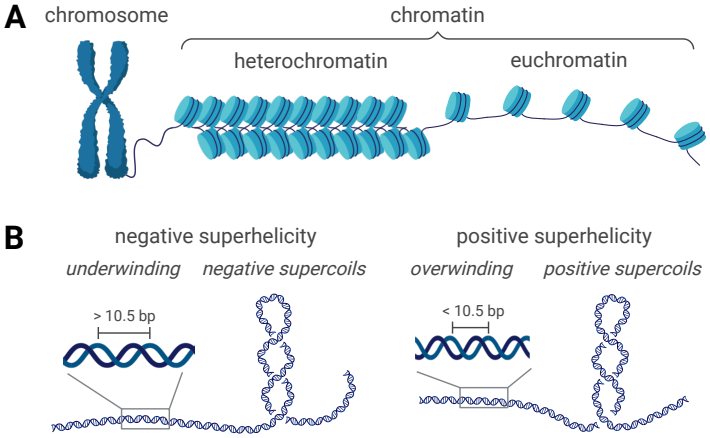


Figure 1.3: Genome accessibility and supercoiling. **A** Nucleosomes (light blue) bind DNA (dark blue) to form chromatin, either tightly packed in heterochromatin or more loosely packed in euchromatin. **B** Negative and positive DNA superhelicity causes under- and overwinding of the DNA helix, and induces supercoil formation.

1.5.3. Genome accessibility and supercoiling

Cas9 activity on a DNA site is limited by the accessibility of the genome at that locus. In prokaryotes, the genome is largely available for protein binding, however in eukaryotes, much of the genome is not. The structure of the eukaryotic genome is precisely regulated in order to control gene expression and to facilitate mitosis, and correspondingly, it varies considerably across cell types and throughout the cell cycle [100, 101]. The chromosomes in a nucleus are present as chromatin, DNA in complex with nucleosomes and other architectural proteins (**Figure 1.3A**). Chromatin can attain different compaction levels; during interphase (outside mitosis), most of the genome consists of less compacted and relatively accessible euchromatin (92% in humans [102]) while some regions are strongly compacted and inaccessible heterochromatin. The accessibility of euchromatin follows from how the nucleosomes wrap DNA, a dynamic process which involves nucleosome breathing (temporary partial DNA unwrapping), and ongoing nucleosome repositioning [101, 103]. In interphase human cells, proteins can bind to 2–3% of the full genome sequence [104].

The protein-accessible bare DNA regions of the genome are not always present as a relaxed, linear helix that rotates every 10.5 bp; often they are overwound or underwound (**Figure 1.3B**). This over- and underwinding has significant impact on Cas9 dynamics, which opens and interrogates DNA faster when underwound and slower when overwound [105, 106]. DNA overwinding and underwinding are described as positive and negative superhelicity, respectively (rather than super- and subhelicity), or DNA torsion. Superhelicity causes a change in the DNA helical pitch as well as the emergence of loops in its secondary structure, known as DNA supercoils. DNA plasmids absorb 70% of superhelicity in supercoils and 30% in helix twist [107], but in linear

DNA this split-up depends on tension [108, 109], and at extreme superhelical levels or tensions, DNA strands may even separate or attain a different structure such as the left-handed Z-form. DNA superhelicity and geometry are also affected by DNA-binding proteins, for instance nucleosomes, each of which wraps DNA in effectively one negative supercoil [103].

Cells employ various proteins that control superhelicity (nucleosomes, topoisomerases [110], and SMCs [111]) in order to regulate expression and allow DNA packaging. Prokaryotes continuously maintain a supercoiled circular genome (negative for bacteria, negative or positive for archaea), while eukaryotes organize their genome in supercoiling domains (topologically associated domains) with distinct superhelical levels [110, 111]. Negative superhelicity is associated with increased gene expression because RNA polymerases can open and transcribe DNA faster when it is underwound. In turn, as polymerases move straight along DNA during transcription, they overwind the DNA ahead and underwind the DNA behind. This example of transcription illustrates that superhelicity levels are in general highly dynamic, specific to the genome context, and coupled to chromatin accessibility. Together, all these aspects make it challenging to systematically measure local DNA state throughout the genome and across cells and cell cycles [112, 113].

1.6. gRNA design: measuring and predicting activity

In order to choose a gRNA or design other aspects of CRISPR applications, it is important to be able to assess how these choices determine the editing outcomes in their cellular context. To do so, experimental platforms have been developed to measure Cas9 activity, and models have been developed that predict activity.

1.6.1. Measuring activity *in vitro* and *in vivo*

There is a wide variety of experimental assays that register DSBs or the editing outcomes due to a CRISPR application. Specifically, rich data is available from high-throughput methods or cleavage “screens”, which either measure the on-target efficiency of many gRNAs in parallel, or determine the specificity of one or few gRNAs by measuring cleavage of all genomic off-targets in parallel.

In vitro methods isolate Cas9 and the genomic DNA in a cell-free setting, and hence reveal Cas9’s intrinsic cleavage dynamics. Compared to *in vivo* screens in cells, these are relatively simple, give fast results, and have a high sensitivity to DSBs, making them particularly useful to detect genome-wide off-target DSBs. Popular assays include Digenome-seq [114], CIRCLE-seq [115], SITE-seq [116], and CHANGE-seq [117]. These approaches rely on next-generation sequencing (NGS) to identify the DSB sites; some by tagging the DSBs with DNA markers, others by aligning the cleavage products with a reference genome.

In vivo methods measure activity in cells and hence report not only on Cas9’s internal target specificity but also on external factors such as chromatin state and DNA repair. To measure on-target efficiency *in vivo*, two approaches can be taken [118]. The first approach reports on functional efficiency, e.g. if a gRNA causes phenotypic change

in terms of for instance gene expression, cell survival or colony growth [119]. This approach can also be taken to quantify the efficiency of CRISPRi/CRISPRa assays in which the genome sequence remains intact. The second approach reports on how efficiently the gene sequence is edited, typically by selectively sequencing the target site with amplicon NGS and detecting the edits that were induced by DNA repair [80]. High-throughput platforms are available for both approaches [118]. To measure gRNA specificity *in vivo*, one can screen the whole genome for DSBs with techniques such as GUIDE-seq [120], BLISS [121], and DISCOVER-seq [122]. These methods use NGS to either directly detect DSBs, to detect target edits induced by DNA repair, or to detect repair activity in terms of DNA-repair-factor binding. While *in vitro* off-target screens identify all DSBs that could possibly occur on a genome, *in vivo* screens reveal which of those are likely to emerge in a cell-specific context.

1.6.2. Predicting activity *in silico*

Often, one wants to select a promising gRNA for gene editing purposes without carrying out multiple laborious whole-genome screens. To do so, numerous *in silico* models are available that predict Cas9's cleavage characteristics based on large *in vivo* cleavage screens. These models provide either on-target scores, estimating efficiency on a selected target, or off-target scores, estimating the specificity [123]. Most end users access these models through online gRNA design tools such as CRISPOR [124], CHOPCHOP [125], and commercial platforms, where users specify a target gene and receive a ranked list of multiple candidate gRNAs together with their on- and off-target metrics.

Early algorithms were based on sequence alignment or heuristic scoring rules [80], but current state-of-the-art algorithms are based on machine learning (ML) techniques [118]. Among the ML models, the older and more well-established ones use conventional ML methods, but more recently there has been a rapid growth of deep learning models [126].

Conventional ML models are trained to features that are manually curated. These features typically include an encoding of the DNA sequence around the 20-nt target, but may also contain additional features such as specific motif presence, GC content or gRNA stability [127]. Popular conventional ML models include Azimuth or "Rule Set 2" (on-target scores) [119], CRISPRscan (on-target scores) [128], and Elevation (off-target scores) [129]. These conventional ML models, alongside heuristic rules and alignment-based scoring schemes, underlie the on- and off-target scores that are presented on most online gRNA design platforms.

Deep learning models use large neural networks to infer Cas9 activity patterns. These models require no feature curation as they learn meaningful features directly from the encoded DNA sequence during training [126]. Deep learning models can capture many details in extensive datasets, however their training is computationally demanding and their complexity makes them hard to interpret [118]. Current widespread models include DeepCRISPR (on- and off-target scores) [130] and DeepSpCas9 (on-target scores) [131]; however, the field of deep-learning-based CRISPR activity prediction is evolving rapidly, with new models being released frequently [126].

The performance of ML models relies heavily on their training data. For instance,

rich data is available for wild-type Cas9 activity in human cells, so there are many different human-specific models to assist in genome editing applications. However, ML models poorly generalize to conditions other than those at which their training data were obtained. To overcome this issue, some models train on multiple data sources simultaneously, for instance across species [129, 130]. Still, repeated benchmarking studies show that when algorithms are applied to new cell types, organisms, CRISPR enzymes, and experimental settings, they rapidly lose performance [118, 123, 124, 127, 132]. Recent work explored the potential of transfer learning to make more generalizable predictions [133, 134]. In this approach, one first trains a base ML model on *in vitro* cleavage data to predict cell-independent activity, and then embeds the base model in a new neural network, which is trained to smaller, context-specific *in vivo* cleavage data.

1.6.3. Biophysical modeling

Biophysical modeling presents an alternative approach to understanding and predicting CRISPR activity. These models typically start from a mechanistic description of Cas9 function, often informed by single-molecule studies. Then, using kinetics and statistical physics, they formulate a mathematical description that captures experimental observations and generalizes to other conditions. Biophysical models have addressed Cas9's target search [135–138], building upon decades of extensive theoretical work on DNA target search by transcription factors [139–141], as well as Cas9's target recognition [91, 133, 142–147], building upon studies of nucleic acid energetics and kinetics [148–154].

Physics-based models have several characteristic strengths compared to the sequence-based, heuristic, and machine-learning models listed in **Section 1.6.2**. First, they are interpretable. Biophysical models are built with physically or mechanistically meaningful parameters, such as binding rates or folding energies. Fitting these parameter values to experimental data reveals the underlying physical properties of the enzyme and hence generates biological insight. Second, they are adaptable. Interpretable parameters allow for adaptation to different conditions. For example, a physics-based model of target binding at low RNP concentrations can be straightforwardly adapted to high RNP concentrations, using knowledge of basic physical chemistry that is typically unavailable in machine learning. Biophysical models can thus successfully generalize to other contexts, provided that it is understood which physical processes modulate the model. Third, they are universal. The dichotomy between on- and off-target activity prediction does not reflect that Cas9 operates on any DNA target according to the same principles. As biophysical models directly address these universal mechanisms of Cas9 activity, they can predict on- and off-target activity in a single framework. Fourth, they are quantitative. In general, the science of physics is about making quantitative predictions. Likewise, biophysical Cas9 models predict concrete and interpretable quantities, such as the probability to cleave a target within the experiment duration, in clear contrast to the on- and off-target scores produced by other activity prediction models. With such interpretable and quantitative output, new calculations can be done, for instance to quantify how the individual risks of off-target DSBs accumulate to overall off-target activity.

1.7. This thesis

This thesis uses biophysical modeling to connect the molecular mechanisms of Cas9 target recognition to its cellular editing outcomes.

Chapter 2 provides theoretical groundwork to connect the molecular and the cellular scale. Cas9's target recognition mechanism is best characterized in *in vitro* experiments, and accordingly, these findings are most informative for Cas9 model making. However, *in vitro* dynamics do not translate straightforwardly to *in vivo* conditions in cells. The chapter presents a kinetic framework that unifies these two regimes, thus laying out the general landscape in which our later chapters are located.

Chapter 3 introduces the model CRISPRzip, which provides a biophysical description of Cas9 target interrogation as a function of gRNA and DNA nucleotide sequence. Building upon a previous kinetic framework from Eslami-Mossallam et al. (2022) [147], CRISPRzip considers the energetics of R-loop formation at base-pair resolution, and hence captures how binding and cleaving kinetics vary across DNA off-targets and gRNAs. In addition to this characterization of intrinsic Cas9 dynamics, the chapter addresses two extrinsic variables that are essential in cells: RNP concentration and DNA superhelicity. The model integrates both variables, and shows that both cause heterogeneous modulation of cleavage dynamics across DNA targets. CRISPRzip exemplifies adaptability in biophysical modeling, as it readily supports general DNA and gRNA sequence, RNP concentration and DNA superhelicity, and allows adaptation to other RNA-guided nucleases thanks to its explicit separation of protein and nucleic acid contributions to R-loop dynamics. The later chapters of this thesis make repeated use of CRISPRzip, adapting it to various cellular conditions.

Chapter 4 more closely investigates the combined effects of nucleic acid sequence and superhelicity on Cas9 target interrogation. The chapter presents an experimental approach to high-throughput profiling of Cas9 cleavage of negatively supercoiled DNA. As such, it is the first report of supercoiling effects on Cas9 activity across thousands of DNA sequences, revealing a heterogeneous response among targets. CRISPRzip is employed to capture and rationalize the observed supercoiling effects. All in all, the chapter spotlights DNA supercoiling as a crucial determinant of Cas9's cellular specificity.

Chapter 5 considers Cas9 activity across a full genome *in vivo*. Drawing on the kinetic framework from the beginning of this thesis and on CRISPRzip, it predicts general cleavage patterns that are independent of organism-specific nucleotide contents. The chapter addresses several effects by which the genome context could modulate the expected cleavage dynamics (Cas9 trapping on non-target sites, competitive cleavage dynamics, background off-target cleavage), each of which strongly depends on DNA superhelical state. The calculations show that none of these genome context effects should affect Cas9 activity in typical cellular settings. This finding supports that Cas9 dynamics can be characterized *in vitro* to predict activity *in vivo*. Finally, a comparison between delivery strategies introduces an abstract quantity, the cleavage capacity, that unifies their different editing outcomes, and can assist in the optimization of efficient and precise CRISPR application design.

Chapter 6 summarizes how all the above findings together pave the way from molecular characterization of CRISPR proteins to predicting genome engineering

success. It concludes the thesis by looking ahead, highlighting relevant knowledge gaps and mapping out the next steps towards physics-based and context-aware CRISPR activity prediction.

References

1. Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* **578,** 229–236. doi:10.1038/s41586-020-1978-5 (2020).
2. Urnov, F. D. Genome Editing B.C. (Before CRISPR): Lasting Lessons from the "Old Testament". *The CRISPR Journal* **1,** 34–46. doi:10.1089/crispr.2018.29007.fyu (2018).
3. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nature Communications* **9,** 1911. doi:10.1038/s41467-018-04252-2 (2018).
4. Lander, E. S. The Heroes of CRISPR. *Cell* **164,** 18–28. doi:10.1016/j.cell.2015.12.041 (2016).
5. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research* **41,** 4360–4377. doi:10.1093/nar/gkt157 (2013).
6. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology* **18.** Number: 2, 67–83. doi:10.1038/s41579-019-0299-x (2020).
7. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakatura, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isoenzyme conversion in Escherichia coli, and identification of the gene product. *Journal of Bacteriology* **169,** 5429–5433. doi:10.1128/jb.169.12.5429-5433.1987 (1987).
8. Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164,** 29–44. doi:10.1016/j.cell.2015.12.035 (2016).
9. Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155,** 733–740. doi:10.1099/mic.0.023960-0 (2009).
10. Leenay, R. T. et al. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell* **62,** 137–147. doi:10.1016/j.molcel.2016.02.031 (2016).
11. Bhattacharya, S. & Satpati, P. Insights into the Mechanism of CRISPR/Cas9-Based Genome Editing from Molecular Dynamics Simulations. *ACS Omega* **8,** 1817–1837. doi:10.1021/acsomega.2c05583 (2023).
12. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821. doi:10.1126/science.1225829 (2012).
13. Nishimasu, H. et al. Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* **156,** 935–949. doi:10.1016/j.cell.2014.02.001 (2014).

14. Cofsky, J. C., Soczek, K. M., Knott, G. J., Nogales, E. & Doudna, J. A. CRISPR–Cas9 bends and twists DNA to read its sequence. *Nature Structural & Molecular Biology* **29,** 395–402. doi:10.1038/s41594-022-00756-0 (2022).

15. Globyte, V., Lee, S. H., Bae, T., Kim, J.-S. & Joo, C. CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. *The EMBO Journal* **38.** doi:10.15252/embj.201899466 (2019).

16. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573. doi:10.1038/nature13579 (2014).

17. Jiang, F. et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351,** 867–871. doi:10.1126/science.aad8282 (2016).

18. Pacesa, M. et al. R-loop formation and conformational activation mechanisms of Cas9. *Nature* **609.** Number: 7925, 191–196. doi:10.1038/s41586-022-05114-0 (2022).

19. Palermo, G. et al. Key role of the REC lobe during CRISPR–Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain. *Quarterly Reviews of Biophysics* **51,** e9. doi:10.1017/S0033583518000070 (2018).

20. Zhu, X. et al. Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nature Structural & Molecular Biology* **26,** 679–685. doi:10.1038/s41594-019-0258-2 (2019).

21. Pacesa, M., Pelea, O. & Jinek, M. Past, present, and future of CRISPR genome editing technologies. *Cell* **187,** 1076–1100. doi:10.1016/j.cell.2024.01.042 (2024).

22. Xue, C. & Greene, E. C. DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing. *Trends in Genetics* **37,** 639–656. doi:10.1016/j.tig.2021.02.008 (2021).

23. Nambiar, T. S., Baudrier, L., Billon, P. & Ciccia, A. CRISPR-based genome editing through the lens of DNA repair. *Molecular Cell* **82,** 348–388. doi:10.1016/j.molcel.2021.12.026 (2022).

24. Song, B., Yang, S., Hwang, G.-H., Yu, J. & Bae, S. Analysis of NHEJ-Based DNA Repair after CRISPR-Mediated DNA Cleavage. *International Journal of Molecular Sciences* **22.** Number: 12, 6397. doi:10.3390/ijms22126397 (2021).

25. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nature Reviews Genetics* **16,** 299–311. doi:10.1038/nrg3899 (2015).

26. Hart, T. et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Genes|Genomes|Genetics* **7,** 2719–2727. doi:10.1534/g3.117.041277 (2017).

27. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews Genetics* **19,** 67–80. doi:10.1038/nrg.2017.97 (2018).

28. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167,** 1853–1866.e17. doi:10.1016/j.cell.2016.11.038 (2016).

29. Du, J. et al. Quantitative assessment of HR and NHEJ activities via CRISPR/Cas9-induced oligodeoxynucleotide-mediated DSB repair. *DNA Repair* **70,** 67–71. doi:10.1016/j.dnarep.2018.09.002 (2018).

30. Qi, L. S. et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152,** 1173–1183. doi:10.1016/j.cell.2013.02.022 (2013).

31. Gilbert, L. A. et al. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154,** 442–451. doi:10.1016/j.cell.2013.06.044 (2013).

32. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533,** 420–424. doi:10.1038/nature17946 (2016).

33. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576,** 149–157. doi:10.1038/s41586-019-1711-4 (2019).

34. Karvelis, T. et al. Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599,** 692–696. doi:10.1038/s41586-021-04058-1 (2021).

35. Saito, M. et al. Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature* **620.** Number: 7974, 660–668. doi:10.1038/s41586-023-06356-2 (2023).

36. Jiang, K. et al. Programmable RNA-guided DNA endonucleases are widespread in eukaryotes and their viruses. *Science Advances* **9,** eadk0171. doi:10.1126/sciadv.adk0171 (2023).

37. Bot, J. F., van der Oost, J. & Geijsen, N. The double life of CRISPR–Cas13. *Current Opinion in Biotechnology* **78,** 102789. doi:10.1016/j.copbio.2022.102789 (2022).

38. Hutvagner, G. & Simard, M. J. Argonaute proteins: Key players in RNA silencing. *Nature Reviews Molecular Cell Biology* **9,** 22–32. doi:10.1038/nrm2321 (2008).

39. Huang, X., Yang, D., Zhang, J., Xu, J. & Chen, Y. E. Recent Advances in Improving Gene-Editing Specificity through CRISPR–Cas9 Nuclease Engineering. *Cells* **11.** Number: 14, 2186. doi:10.3390/cells11142186 (2022).

40. Kleinstiver, B. P. et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495. doi:10.1038/nature16526 (2016).

41. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351,** 84–88. doi:10.1126/science.aad5227 (2016).

42. Chen, J. S. et al. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550,** 407–410. doi:10.1038/nature24268 (2017).

43. Riesenberg, S., Helmbrecht, N., Kanis, P., Maricic, T. & Pääbo, S. Improved gRNA secondary structures allow editing of target sites resistant to CRISPR-Cas9 cleavage. *Nature Communications* **13.** doi:10.1038/s41467-022-28137-7 (2022).

44. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368,** 290–296. doi:10.1126/science.aba8853 (2020).

45. Silverstein, R. A. et al. Custom CRISPR–Cas9 PAM variants via scalable engineering and machine learning. *Nature* **643,** 539–550. doi:10.1038/s41586-025-09021-y (2025).

46. Nishida, K. & Kondo, A. CRISPR-derived genome editing technologies for metabolic engineering. *Metabolic Engineering. Tools and Strategies of Metabolic Engineering* **63,** 141–147. doi:10.1016/j.ymben.2020.12.002 (2021).

47. Li, Y. et al. Recent advances in high-throughput metabolic engineering: Generation of oligonucleotide-mediated genetic libraries. *Biotechnology Advances* **59,** 107970. doi:10.1016/j.biotechadv.2022.107970 (2022).

48. Burbano, D. A. et al. CRISPR Tools for Engineering Prokaryotic Systems: Recent Advances and New Applications. *Annual Review of Chemical and Biomolecular Engineering* **15,** 389–430. doi:10.1146/annurev-chembioeng-100522-114706 (2024).

49. Glinšek, K., Bozovičar, K. & Bratkovič, T. CRISPR Technologies in Chinese Hamster Ovary Cell Line Engineering. *International Journal of Molecular Sciences* **24,** 8144. doi:10.3390/ijms24098144 (2023).

50. Garg, D. et al. Revolutionizing biofuel generation: Unleashing the power of CRISPR-Cas mediated gene editing of extremophiles. *Microbiological Research* **274,** 127443. doi:10.1016/j.micres.2023.127443 (2023).

51. Dudeja, C., Mishra, A., Ali, A., Singh, P. P. & Jaiswal, A. K. Microbial Genome Editing with CRISPR–Cas9: Recent Advances and Emerging Applications Across Sectors. *Fermentation* **11,** 410. doi:10.3390/fermentation11070410 (2025).

52. Jiang, W. et al. Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Research* **41,** e188. doi:10.1093/nar/gkt780 (2013).

53. Tuncel, A., Pan, C., Clem, J. S., Liu, D. & Qi, Y. CRISPR–Cas applications in agriculture and plant research. *Nature Reviews Molecular Cell Biology* **26,** 419–441. doi:10.1038/s41580-025-00834-3 (2025).

54. Zhu, H., Li, C. & Gao, C. Applications of CRISPR–Cas in agriculture and plant biotechnology. *Nature Reviews Molecular Cell Biology* **21,** 661–677. doi:10.1038/s41580-020-00288-9 (2020).

55. Li, T. et al. Domestication of wild tomato is accelerated by genome editing. *Nature Biotechnology* **36,** 1160–1163. doi:10.1038/nbt.4273 (2018).

56. Genetic Literacy Project. *Global Gene Editing Regulation Tracker* https://crispr-gene-editing-regs-tracker.geneticliteracyproject.org. 2024.

57. De Lange, J., Nalley, L. L., Yang, W., Shew, A. & de Steur, H. The future of CRISPR gene editing according to plant scientists. *iScience* **25.** doi:10.1016/j.isci.2022.105012 (2022).

58. Bassett, A. R. Editing the genome of hiPSC with CRISPR/Cas9: disease models. *Mammalian Genome* **28,** 348–364. doi:10.1007/s00335-017-9684-9 (2017).

59. Geurts, M. H. & Clevers, H. CRISPR engineering in organoids for gene repair and disease modelling. *Nature Reviews Bioengineering* **1,** 32–45. doi:10.1038/s44222-022-00013-5 (2023).

60. Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360,** 436–439. doi:10.1126/science.aar6245 (2018).

61. Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356,** 438–442. doi:10.1126/science.aam9321 (2017).

62. Li, L. et al. CRISPR-Cas-mediated diagnostics. *Trends in Biotechnology* **40,** 1326–1345. doi:10.1016/j.tibtech.2022.04.006 (2022).

63. Kyrou, K. et al. A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes. *Nature Biotechnology* **36,** 1062–1066. doi:10.1038/nbt.4245 (2018).

64. Morshedzadeh, F. et al. An Update on the Application of CRISPR Technology in Clinical Practice. *Molecular Biotechnology* **66,** 179–197. doi:10.1007/s12033-023-00724-z (2024).

65. Frangoul, H. et al. CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *New England Journal of Medicine* **384,** 252–260. doi:10.1056/NEJMoa2031054 (2021).

66. Gillmore, J. D. et al. CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. *New England Journal of Medicine* **385,** 493–502. doi:10.1056/NEJMoa2107454 (2021).

67. Cohn, D. M. et al. CRISPR-Based Therapy for Hereditary Angioedema. *New England Journal of Medicine* **392,** 458–467. doi:10.1056/NEJMoa2405734 (2025).

68. Pierce, E. A. et al. Gene Editing for CEP290-Associated Retinal Degeneration. *New England Journal of Medicine* **390,** 1972–1984. doi:10.1056/NEJMoa2309915 (2024).

69. Chehelgerdi, M. et al. Comprehensive review of CRISPR-based gene editing: mechanisms, challenges, and applications in cancer therapy. *Molecular Cancer* **23,** 9. doi:10.1186/s12943-023-01925-5 (2024).

70. Stadtmauer, E. A. et al. CRISPR-engineered T cells in patients with refractory cancer. *Science* **367,** eaba7365. doi:10.1126/science.aba7365 (2020).

71. Xu, L. et al. CRISPR-Edited Stem Cells in a Patient with HIV and Acute Lymphocytic Leukemia. *New England Journal of Medicine* **381,** 1240–1247. doi:10.1056/NEJMoa1817426 (2019).

72. Abbott, T. R. et al. Development of CRISPR as an Antiviral Strategy to Combat SARS-CoV-2 and Influenza. *Cell* **181,** 865–876.e12. doi:10.1016/j.cell.2020.04.020 (2020).

73. Zahedipour, F., Zahedipour, F., Zamani, P., Jaafari, M. R. & Sahebkar, A. Harnessing CRISPR technology for viral therapeutics and vaccines: from preclinical studies to clinical applications. *Virus Research* **341,** 199314. doi:10.1016/j.virusres.2024.199314 (2024).

74. Jha, A. et al. Systematic Review: Gene Editing for Cardiovascular Disease. *Journal of Heart Valve Disease* **30,** 160–166 (2025).

75. Committee on Human Gene Editing: Scientific, Medical, and Ethical Considerations, National Academy of Sciences, National Academy of Medicine & National Academies of Sciences, Engineering, and Medicine. *Human Genome Editing: Science, Ethics, and Governance* Pages: 24623. ISBN: 978-0-309-45288-5. doi:10.17226/24623 (National Academies Press, Washington, D.C., 2017).

76. Lanphier, E., Urnov, F., Haecker, S. E., Werner, M. & Smolenski, J. Don't edit the human germ line. *Nature* **519,** 410–411. doi:10.1038/519410a (2015).

77. Liang, P. et al. CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein & Cell* **6,** 363–372. doi:10.1007/s13238-015-0153-5 (2015).

78. Ma, H. et al. Correction of a pathogenic gene mutation in human embryos. *Nature* **548,** 413–419. doi:10.1038/nature23305 (2017).

79. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology* **31,** 822–827. doi:10.1038/nbt.2623 (2013).

80. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* **31,** 827–832. doi:10.1038/nbt.2647 (2013).

81. Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular Therapy - Nucleic Acids* **4,** e264. doi:10.1038/mtna.2015.37 (2015).

82. Guo, C., Ma, X., Gao, F. & Guo, Y. Off-target effects in CRISPR/Cas9 gene editing. *Frontiers in Bioengineering and Biotechnology* **11.** doi:10.3389/fbioe.2023.1143157 (2023).

83. Jiang, L. et al. CRISPR/Cas9-Induced DNA Damage Enriches for Mutations in a p53-Linked Interactome: Implications for CRISPR-Based Therapies. *Cancer Research* **82,** 36–45. doi:10.1158/0008-5472.CAN-21-1692 (2022).

84. Tycko, J. et al. Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nature Communications* **10,** 4063. doi:10.1038/s41467-019-11955-7 (2019).

85. Rostain, W. et al. Cas9 off-target binding to the promoter of bacterial genes leads to silencing and toxicity. *Nucleic Acids Research* **51,** 3485–3496. doi:10.1093/nar/gkad170 (2023).

86. Rohatgi, N. et al. Seed sequences mediate off-target activity in the CRISPR-interference system. *Cell Genomics* **4.** doi:10.1016/j.xgen.2024.100693 (2024).

87. Höijer, I. et al. CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo that segregate across generations. *Nature Communications* **13,** 627. doi:10.1038/s41467-022-28244-5 (2022).

88. Adikusuma, F. et al. Large deletions induced by Cas9 cleavage. *Nature* **560,** E8–E9. doi:10.1038/s41586-018-0380-z (2018).

89. Cullot, G. et al. CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nature Communications* **10,** 1–14. doi:10.1038/s41467-019-09006-2 (2019).

90. Boyle, E. A. et al. Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Science Advances* **7.** doi:10.1126/sciadv.abe5496 (2021).

91. Zhang, D., Hurst, T., Duan, D. & Chen, S. J. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 8693–8698. doi:10.1073/pnas.1820523116 (2019).

92. Raguram, A., Banskota, S. & Liu, D. R. Therapeutic *in vivo* delivery of gene editing agents. *Cell* **185,** 2806–2827. doi:10.1016/j.cell.2022.03.045 (2022).

93. Severi, A. A. & Akbari, B. CRISPR-Cas9 delivery strategies and applications: Review and update. *genesis* **62,** e23598. doi:10.1002/dvg.23598 (2024).

94. Karp, H. et al. Packaged delivery of CRISPR–Cas9 ribonucleoproteins accelerates genome editing. *Nucleic Acids Research* **53,** gkaf105. doi:10.1093/nar/gkaf105 (2025).

95. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology* **20,** 698–714. doi:10.1038/s41580-019-0152-0 (2019).

96. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature Biotechnology* **34,** 339–344. doi:10.1038/nbt.3481 (2016).

97. Maruyama, T. et al. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nature Biotechnology* **33,** 538–542. doi:10.1038/nbt.3190 (2015).

98. Chu, V. T. et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature Biotechnology* **33,** 543–548. doi:10.1038/nbt.3198 (2015).

99. Liu, M. S. et al. Engineered CRISPR/Cas9 enzymes improve discrimination by slowing DNA cleavage to allow release of off-target DNA. *Nature Communications* **11.** doi:10.1038/s41467-020-17411-1 (2020).

100. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82. doi:10.1038/nature11232 (2012).

101. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20,** 207–220. doi:10.1038/s41576-018-0089-8 (2019).

102. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945. doi:10.1038/nature03001 (2004).

103. McGinty, R. K. & Tan, S. Nucleosome Structure and Function. *Chemical Reviews* **115,** 2255–2273. doi:10.1021/cr500373h (2015).

104. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82. doi:10.1038/nature11232 (2012).

105. Ivanov, I. E. et al. Cas9 interrogates DNA in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 5853–5860. doi:10.1073/pnas.1913445117 (2020).

106. Newton, M. D. et al. Negative DNA supercoiling induces genome-wide Cas9 off-target activity. *Molecular Cell* **83,** 3533–3545.e5. doi:10.1016/j.molcel.2023.09.008 (2023).

107. Boles, T. C., White, J. H. & Cozzarelli, N. R. Structure of plectonemically supercoiled DNA. *Journal of Molecular Biology* **213,** 931–951. doi:10.1016/S0022-2836(05)80272-4 (1990).

108. Marko, J. F. in *Mathematics of DNA Structure, Function and Interactions* 1st ed. *The IMA Volumes in Mathematics and its Applications* 150, 225–249 (Springer, New York, NY, 2020). ISBN: 978-1-4419-0669-4.

109. Gao, X., Hong, Y., Ye, F., Inman, J. T. & Wang, M. D. Torsional Stiffness of Extended and Plectonemic DNA. *Physical Review Letters* **127,** 028101. doi:10.1103/PhysRevLett.127.028101 (2021).

110. Duprey, A. & Groisman, E. A. The regulation of DNA supercoiling across evolution. *Protein Science* **30,** 2042–2056. doi:10.1002/pro.4171 (2021).

111. Gilbert, N. & Allan, J. Supercoiling in DNA and chromatin. *Current Opinion in Genetics & Development. Genome architecture and expression* **25,** 15–21. doi:10.1016/j.gde.2013.10.013 (2014).

112. Naughton, C. et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nature Structural & Molecular Biology* **20,** 387–395. doi:10.1038/nsmb.2509 (2013).

113. Yao, Q., Zhu, L., Shi, Z., Banerjee, S. & Chen, C. Topoisomerase-modulated genome-wide DNA supercoiling domains colocalize with nuclear compartments and regulate human gene expression. *Nature Structural & Molecular Biology* **32,** 48–61. doi:10.1038/s41594-024-01377-5 (2025).

114. Kim, D. et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature Methods* **12,** 237–243. doi:10.1038/nmeth.3284 (2015).

115. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nature Methods* **14,** 607–614. doi:10.1038/nmeth.4278 (2017).

116. Cameron, P. et al. Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nature Methods* **14,** 600–606. doi:10.1038/nmeth.4284 (2017).

117. Lazzarotto, C. R. et al. CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity. *Nature Biotechnology* **38,** 1317–1327. doi:10.1038/s41587-020-0555-7 (2020).

118. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Research* **50,** 3616–3637. doi:10.1093/nar/gkac192 (2022).

119. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34,** 184–191. doi:10.1038/nbt.3437 (2016).

120. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology* **33,** 187–197. doi:10.1038/nbt.3117 (2015).

121. Yan, W. X. et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications* **8,** 15058. doi:10.1038/ncomms15058 (2017).

122. Wienert, B. et al. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science* **364,** 286–289. doi:10.1126/science.aav9023 (2019).

123. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology* **17,** 148. doi:10.1186/s13059-016-1012-2 (2016).

124. Concordet, J.-P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research* **46,** W242–W245. doi:10.1093/nar/gky354 (2018).

125. Labun, K. et al. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research* **47,** W171–W174. doi:10.1093/nar/gkz365 (2019).

126. Sherkatghanad, Z., Abdar, M., Charlier, J. & Makarenkov, V. Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review. *Briefings in Bioinformatics* **24,** bbad131. doi:10.1093/bib/bbad131 (2023).

127. Zhang, G., Luo, Y., Dai, X. & Dai, Z. Benchmarking deep learning methods for predicting CRISPR/Cas9 sgRNA on- and off-target activities. *Briefings in Bioinformatics* **24,** bbad333. doi:10.1093/bib/bbad333 (2023).

128. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods* **12,** 982–988. doi:10.1038/nmeth.3543 (2015).

129. Listgarten, J. et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering* **2,** 38–47. doi:10.1038/s41551-017-0178-6 (2018).

130. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biology* **19,** 80. doi:10.1186/s13059-018-1459-4 (2018).

131. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. *Science Advances* **5,** eaax9249. doi:10.1126/sciadv.aax9249 (2019).

132. Yuan, H. et al. An Overview and Comparative Analysis of CRISPR-SpCas9 gRNA Activity Prediction Tools. *The CRISPR Journal* **8,** 89–104. doi:10.1089/crispr.2024.0058 (2025).

133. Zhang, Z., Lamson, A. R., Shelley, M. & Troyanskaya, O. Interpretable neural architecture search and transfer learning for understanding CRISPR–Cas9 off-target enzymatic reactions. *Nature Computational Science* **3,** 1056–1066. doi:10.1038/s43588-023-00569-1 (2023).

134. Elkayam, S., Tziony, I. & Orenstein, Y. DeepCRISTL: deep transfer learning to predict CRISPR/Cas9 on-target editing efficiency in specific cellular contexts. *Bioinformatics* **40,** btae481. doi:10.1093/bioinformatics/btae481 (2024).

135. Shvets, A. A. & Kolomeisky, A. B. Mechanism of Genome Interrogation: How CRISPR RNA-Guided Cas9 Proteins Locate Specific Targets on DNA. *Biophysical Journal* **113,** 1416–1424. doi:10.1016/j.bpj.2017.08.013 (2017).

136. Thonnekottu, D. & Chatterjee, D. CRISPR–Cas9 Genome Interrogation: A Facilitated Subdiffusive Target Search Strategy. *The Journal of Physical Chemistry B* **124,** 3271–3282. doi:10.1021/acs.jpcb.0c00086 (2020).

137. Lu, Q., Bhat, D., Stepanenko, D. & Pigolotti, S. Search and Localization Dynamics of the CRISPR-Cas9 System. *Physical Review Letters* **127,** 208102. doi:10.1103/PhysRevLett.127.208102 (2021).

138. Lu, Q. & Pigolotti, S. Target search in the CRISPR/Cas9 system: facilitated diffusion with target cues. *New Journal of Physics* **26,** 043007. doi:10.1088/1367-2630/ad3841 (2024).

139. Berg, O. G. & von Hippel, P. H. Diffusion-controlled macromolecular interactions. *Annual Review of Biophysics and Biophysical Chemistry* **14,** 131–160. doi:10.1146/annurev.bb.14.060185.001023 (1985).

140. Slutsky, M. & Mirny, L. A. Kinetics of protein-DNA interaction: Facilitated target location in sequence-dependent potential. *Biophysical Journal* **87.** ISBN: 00063495/04/12/4, 4021–4035. doi:10.1529/biophysj.104.050765 (2004).

141. Mirny, L. et al. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *Journal of Physics A: Mathematical and Theoretical* **42,** 434013. doi:10.1088/1751-8113/42/43/434013 (2009).

142. Farasat, I. & Salis, H. M. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLOS Computational Biology* **12,** e1004724. doi:10.1371/journal.pcbi.1004724 (2016).

143. Bisaria, N., Jarmoskaite, I. & Herschlag, D. Lessons from Enzyme Kinetics Reveal Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-Based Genome Editing. *Cell Systems* **4,** 21–29. doi:10.1016/j.cels.2016.12.010 (2017).

144. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biology* **19,** 1–13. doi:10.1186/s13059-018-1534-x (2018).

145. Marklund, E. G. et al. Sequence specificity in DNA binding is mainly governed by association. *Science* **375,** 442–445. doi:10.1126/science.abg7427 (2022).

146. Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depken, M. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Reports* **22,** 1413–1423. doi:10.1016/j.celrep.2018.01.045 (2018).

147. Eslami-Mossallam, B. et al. A kinetic model predicts SpCas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity. *Nature Communications* **13.** doi:10.1038/s41467-022-28994-2 (2022).

148. SantaLucia, J. & Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics* **33,** 415–440. doi:10.1146/annurev.biophys.32.110601.141800 (2004).

149. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research* **38,** D280–D282. doi:10.1093/nar/gkp892 (2010).

150. Banerjee, D. et al. Improved nearest-neighbor parameters for the stability of RNA/DNA hybrids under a physiological condition. *Nucleic Acids Research* **48,** 12042–12054. doi:10.1093/nar/gkaa572 (2020).

151. Irmisch, P., Ouldridge, T. E. & Seidel, R. Modeling DNA-Strand Displacement Reactions in the Presence of Base-Pair Mismatches. *Journal of the American Chemical Society* **142,** 11451–11463. doi:10.1021/jacs.0c03105 (2020).

152. Ratajczyk, E. J., Šulc, P., Turberfield, A. J., Doye, J. P. K. & Louis, A. A. Coarse-grained modeling of DNA–RNA hybrids. *The Journal of Chemical Physics* **160,** 115101. doi:10.1063/5.0199558 (2024).

153. Smith, F. G., Goertz, J. P., Jurinović, K., Stevens, M. M. & Ouldridge, T. E. Strong sequence–dependence in RNA/DNA hybrid strand displacement kinetics. *Nanoscale* **16,** 17624–17637. doi:10.1039/D4NR00542B (2024).

154. Ratajczyk, E. J. et al. Controlling DNA–RNA strand displacement kinetics with base distribution. *Proceedings of the National Academy of Sciences* **122,** e2416988122. doi:10.1073/pnas.2416988122 (2025).

2

A general kinetic framework unifies Cas9 applications *in vitro* and *in vivo*

CRISPR-Cas9 cleavage has been studied both in cells and in cell-free assays, but these settings often reveal different kinetics. Here we unify the two regimes with a minimal model of Cas9 kinetics. The model considers heterogeneous R-loop formation among target sites, and expresses PAM binding as a function of PAM and Cas9 presence. We map out two kinetic limits that are useful to describe activity in in vitro *and* in vivo *experiments. Thanks to the theory of Cas9 activity presented here, we can formulate context-specific cleavage rates in the next chapters, and generalize our findings to other experimental conditions.*

Figure 2.1: Simple kinetic model of target cleavage. The 3-state model representing PAM-binding (k_{PAM}), PAM-unbinding (k_{unb}), and R-loop formation and cleavage ($k_{\mathrm{R},i}$) at PAM site i. The rates k_{PAM} and k_{unb} are assumed to be site-independent; the rate $k_{\mathrm{R},i}$ depends on the DNA sequence of site i.

2.1. Introduction

Successful CRISPR-Cas9 applications rely on extensive characterization of its cleavage activity. For instance, numerous *in vivo* Cas9 screens in human cell lines have generated a rich source of data to understand Cas9's dynamics and to develop activity prediction models [1, 2]. These cleavage patterns observed *in vivo* reflect the complexity of the cellular and nuclear environment, in which final editing outcomes depend on Cas9 activity as well as chromatin state, DNA repair, cellular crowding, and more. An opposite approach is provided by cell-free *in vitro* cleavage assays, which offer a high degree of control over the experimental conditions [3–7]. While both *in vivo* and *in vitro* approaches study the same molecular machinery, they often reveal different aspects of Cas9 dynamics and describe them with different theoretical frameworks.

In this chapter, we study a minimal model for Cas9 activity that connects the observed dynamics in *in vitro* and *in vivo* contexts. We formulate how cleavage depends on PAM-binding and -unbinding, on heterogeneous R-loop kinetics, and on the availability of Cas9:gRNA ribonucleoprotein (RNP) and PAM sites. From our general expressions, we recover well-known descriptions of *in vivo* cleavage kinetics in the "rare-binding limit" and *in vitro* cleavage kinetics in the "PAM-equilibrium limit". By charting the full landscape of Cas9 activity and its kinetic regimes, our simplified theoretical framework unifies the interpretation of cleavage measurements outside and inside cells.

2.2. Simple kinetic model of target cleavage

We introduce a simple model of Cas9 activity on a particular DNA site in order to obtain general cleavage kinetics. The model, presented in **Figure 2.1**, comprises three states that represent how DNA is bound, unbound and cleaved. DNA binding sites are located directly upstream of a protospacer adjacent motif (PAM, 5'-NGG-3'), which (*S. pyogenes*) Cas9 requires to initiate target interrogation. The rates k_{PAM} and k_{unb} quantify how fast an individual PAM site gets bound and released. When site i is PAM-bound, Cas9 can form an R-loop and cleave it at a rate $k_{\mathrm{R},i}$, depending on the complementarity between the target sequence of site i and the gRNA. R-loop formation is a complex multi-state process but the resulting cleavage dynamics are effectively exponential [7] (**Chapter 3**), so we represent it with a single effective rate $k_{\mathrm{R},i}$. Together, the PAM and

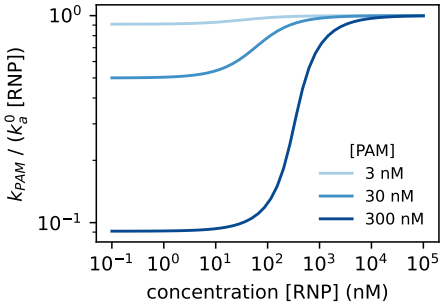


Figure 2.2: Rate of PAM-binding. Context dependence of k_{PAM} according to **Equation 2.2**, with $K_D = 30$ nM.

R-loop dynamics give rise to an overall effective cleavage rate of

$$k_{\text{clv},i} = \frac{k_{\text{PAM}} \cdot k_{\text{R},i}}{k_{\text{PAM}} + k_{\text{unb}} + k_{\text{R},i}}, \tag{2.1}$$

as we derive from first passage time theory in **Suppl. Note 2.S1**. Under typical conditions, cleavage dynamics follow this single rate $k_{\text{clv},i}$, but in a setting where $k_{\text{PAM}} \approx k_{\text{R},i}$ while also $k_{\text{PAM}} \gtrsim k_{\text{unb}}$ and $k_{\text{R},i} \gtrsim k_{\text{unb}}$, one should consider the full cleavage PDF—a hypoexponential distribution with two exponential modes (**Suppl. Note 2.S1**).

Before we simplify **Equation 2.1** to specific *in vitro* and *in vivo* conditions, we specify how the PAM-binding rate k_{PAM} depends on experimental context.

2.3. Rate of PAM-binding

The PAM-binding rate k_{PAM} in a particular setting depends on the abundance of RNPs as well as the number of PAM sites they interact with. We assume that on the timescales of target cleavage ($1/k_{\text{clv},i}$), there is always a local equilibrium between freely diffusing and PAM-bound RNPs. Under this assumption, the PAM-binding rate is the product of the bimolecular association rate k_{a}^0 (units nM^{-1}s^{-1}) and the equilibrium free RNP concentration $[\text{RNP}_{\text{free}}]_{\text{eq}}$. We obtain $[\text{RNP}_{\text{free}}]_{\text{eq}}$ in terms of the total concentrations [RNP] and [PAM] (**Suppl. Note 2.S2**), yielding (**Figure 2.2**)

$$k_{\text{PAM}} = k_{\text{a}}^0 \cdot \frac{1}{2}\left([\text{RNP}] - [\text{PAM}] - K_D + \sqrt{([\text{RNP}] + [\text{PAM}] + K_D)^2 - 4[\text{RNP}][\text{PAM}]}\right). \tag{2.2}$$

Here, $K_D = k_{\text{unb}}/k_{\text{a}}^0$ (in nM) is the dissociation constant between the RNPs and PAM sites. The expression for k_{PAM} is consistent with the assumption of a local RNP equilibrium, as the equilibrium is established faster than any rate $k_{\text{clv},i}$ at which a target i gets cleaved (**Suppl. Note 2.S3**). The expression for k_{PAM} remains valid when the total

concentrations [RNP] and [PAM] evolve with time due to cleavage or other processes that are slower than RNP equilibration.

In our derivation of the free RNP concentration, we have excluded the possibility that Cas9 interacts with DNA beyond PAM binding. In **Chapter 5**, we investigate this assumption by predicting how frequently Cas9 encounters random genomic DNA sequences on which it forms stable partial R-loops. Also, we do not consider the effects of DNA sliding by Cas9, by which it could shuttle between neighboring PAM sites [8]. Cas9 slides along bare DNA in a region of approximately 10 bps [9, 10] and arguably less *in vivo* due to the presence of other DNA-binding proteins. These distances are shorter than the typical 16 bp separation of PAM sites on random DNA, making significant shuttling between neighboring PAM sites unlikely.

The expression for $[\text{RNP}_{\text{free}}]_{\text{eq}}$ simplifies when either the RNPs or PAM sites are relatively abundant (**Suppl. Note 2.S4**), which yields PAM-binding rates

$$k_{\text{PAM}} = \begin{cases} k_{\text{a}}^0\,[\text{RNP}] & \text{if } [\text{RNP}] \gg [\text{PAM}] + K_D, \quad (2.3) \\ k_{\text{a}}^0 \left(\dfrac{K_D}{[\text{PAM}] + K_D} \right) [\text{RNP}] & \text{if } [\text{RNP}] \ll [\text{PAM}] + K_D. \quad (2.4) \end{cases}$$

When much RNP is present, each individual DNA site is bound at a fixed rate. When RNP is scarce, the rate k_{PAM} depends on the genetic context, as high genome densities $[\text{PAM}] \gtrsim K_D$ limit free RNP availability.

In a cellular context, where molecule counts are low, it is more appropriate to consider discrete copy numbers rather than concentrations. Also, single-molecule experiments in cells commonly report the microscopic rate k_{bnd} at which an individual RNP binds any PAM site, in contrast to the rate k_{PAM} at which an individual PAM site i gets bound by any RNP. Therefore, for *in vivo* applications, we express **Equation 2.4** in terms of RNP and PAM copy numbers and k_{bnd}. The copy numbers are defined as $C = [\text{RNP}] \cdot N_A V$ RNP complexes and $N = [\text{PAM}] \cdot N_A V$ PAM sites, with N_A Avogadro's number and V the cellular or nuclear volume. The single-RNP binding rate k_{bnd} is defined as $k_{\text{bnd}} = k_{\text{a}}^0 \cdot [\text{PAM}_{\text{free}}]_{\text{eq}}$, similar to k_{PAM}. Although in general not all PAMs are available for binding ($[\text{PAM}_{\text{free}}]_{\text{eq}} < [\text{PAM}]$), in typical *in vivo* conditions there is a great abundance of PAM sites compared to RNPs, such that $[\text{PAM}_{\text{free}}]_{\text{eq}} \approx [\text{PAM}]$ (**Suppl. Note 2.S4**) and hence $k_{\text{bnd}} \approx k_{\text{a}}^0 \cdot [\text{PAM}] = N k_{\text{a}}^0 / (N_A V)$. In terms of C, N, and k_{bnd}, the PAM-binding rate k_{PAM} in the Cas9-scarce limit becomes

$$k_{\text{PAM}} = \frac{C}{N} \left(\frac{1}{k_{\text{bnd}}} + \frac{1}{k_{\text{unb}}} \right)^{-1} \quad \text{if } C \ll N. \quad (2.5)$$

2.4. Enzymatic limits *in vivo* and *in vitro*

Equation 2.1 of the effective cleavage rate $k_{\text{clv},i}$ simplifies in various kinetic limits (**Figure 2.3**). First, the diffusion-limited and the reaction-limited regimes, two canonical limits in enzyme kinetics. In this system, cleavage becomes diffusion-limited when R-loop formation is very fast, $k_{\text{R},i} \gg k_{\text{unb}} + k_{\text{PAM}}$, such that initial PAM binding sets the effective cleavage rate, $k_{\text{clv},i} = k_{\text{PAM}}$. Cleavage becomes reaction-limited when

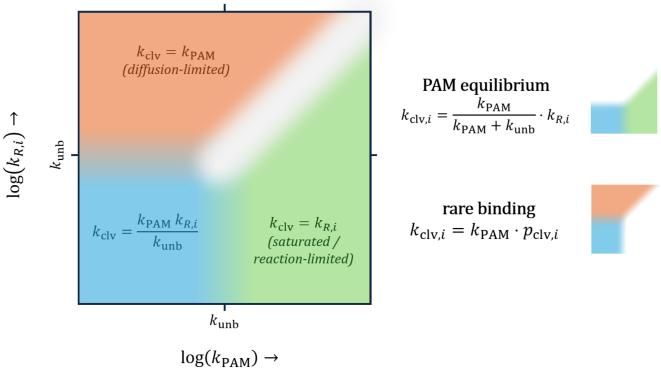


Figure 2.3: Enzymatic limits of Cas9 activity. Illustrative phase diagram for $k_{\text{clv},i}$ according to **Equation 2.1**. In the white region, cleavage dynamics follow a hypoexponential distribution (**Suppl. Note 2.S1**).

PAM association is very fast, $k_{\text{PAM}} \gg k_{\text{unb}} + k_{\text{R},i}$, such that R-loop formation sets the effective cleavage rate, $k_{\text{clv},i} = k_{\text{R},i}$. Below, we discuss how these canonical limits should be extended to describe typical *in vivo* and *in vitro* conditions.

In vivo

In some cells and for some DNA targets, Cas9 activity may be diffusion-limited. For instance, single-molecule tracking experiments suggest diffusion-limited perfect-target cleavage in the context of *E. coli*. One experiment observed $k_{\text{unb}} = 60 \pm 13\,\text{s}^{-1}$ and $k_{\text{bnd}} = 40 \pm 12\,\text{s}^{-1}$ [11], suggesting that a single RNP binds a specific PAM in $1/k_{\text{PAM}} = 11.6 \pm 2.3$ hours on average (**Equation 2.5**, assuming $N = 10^6$ PAM sites [6]). Another experiment measured a similar time for a single dead Cas9 (dCas9) to irreversibly bind the perfect target i, $1/k_{\text{clv},i} = 6.2 \pm 1.4$ hours [6]. Together, these results suggest that an RNP binds the PAM of the perfect target only once before it establishes a stable R-loop, so $k_{\text{R},i} \gg k_{\text{unb}}$. As long as RNP copy numbers are not extremely high ($C \ll 10^6$ so that $k_{\text{PAM}} \ll k_{\text{unb}}$), such direct on-target recognition implies diffusion-limited kinetics. However, diffusion-limited cleavage is not general *in vivo*: direct on-target recognition is not universal, and even when it does apply, off-target sites in the genome have smaller $k_{\text{R},i}$ and are bound repeatedly before cleavage.

The more general and useful limit of **Equation 2.1** to consider in an *in vivo* context is the rare-binding limit (**Figure 2.3**).

$$k_{\text{clv},i} = k_{\text{PAM}} \cdot p_{\text{clv,i}} \quad \text{if } k_{\text{PAM}} \ll k_{\text{unb}} + k_{\text{R},i}, \tag{2.6}$$

where $p_{\text{clv},i}$ corresponds to the probability to cleave upon PAM-binding,

$$p_{\text{clv},i} = \frac{k_{\text{R},i}}{k_{\text{R},i} + k_{\text{unb}}}. \tag{2.7}$$

To explicitly show dependence on RNP count C, PAM count N, and the single-RNP binding rate k_{bnd}, we substitute **Equation 2.5** and obtain

$$k_{\text{clv},i} = \frac{C}{N}\left(\frac{1}{k_{\text{bnd}}} + \frac{1}{k_{\text{unb}}}\right)^{-1} \cdot p_{\text{clv,i}} \quad \text{if } C \ll N \text{ and } C \cdot k_{\text{unb}}\, k_{\text{bnd}} \ll N(k_{\text{bnd}} + k_{\text{unb}})(k_{\text{unb}} + k_{\text{R},i}). \tag{2.8}$$

Equations 2.6 and 2.8 apply to all targets i that are rarely bound by Cas9: their unbound time $1/k_{\text{PAM}}$ should be much longer than their bound time $1/(k_{\text{unb}} + k_{\text{R},i})$. This condition is generally satisfied for all targets in a cell, regardless of $p_{\text{clv},i}$, as RNPs are greatly outnumbered by PAM sites and target search is slow (as illustrated by the above values for k_{bnd}, k_{unb}, and N in *E. coli*). The rare-binding limit excludes reaction-limited kinetics.

The above rare-binding expressions for $k_{\text{clv},i}$ are more universal than the simple 3-state model used in this chapter, so the definition of $p_{\text{clv},i}$ is not limited to **Equation 2.7**. In **Chapter 5**, we consider Cas9 activity *in vivo* using the rare-binding limit, with values for $p_{\text{clv},i}$ derived from a more accurate 21-state model.

In vitro

Reaction-limited kinetics apply to *in vitro* cleavage assays with saturating RNP concentrations, i.e. abundant RNP to ensure that effectively every PAM site i is Cas9-bound. In **Chapters 3 and 4**, such saturating conditions allow us to measure only R-loop kinetics. The precise concentration requirements are $[\text{RNP}] \gg [\text{PAM}] + K_D$ (from **Equation 2.3**) as well as $[\text{RNP}] \gg k_{\text{R},i}/k_{\text{a}}^0$ (from the reaction-limit condition). These conditions are not always guaranteed, as the RNP concentration may be varied [3, 5, 7], or the equilibrium constant K_D may be unknown (**Chapter 3**).

Here, the more general and useful limit of **Equation 2.1** to consider *in vitro* is the PAM-equilibrium limit, which applies when PAM binding and unbinding is fast compared to R-loop formation (**Figure 2.3**),

$$k_{\text{clv},i} = \frac{k_{\text{PAM}}}{k_{\text{PAM}} + k_{\text{unb}}}\, k_{\text{R},i} \quad \text{if } k_{\text{PAM}} + k_{\text{unb}} \gg k_{\text{R},i}. \tag{2.9}$$

The fraction corresponds to the equilibrium occupancy of a particular PAM site i. This PAM-equilibrium limit excludes diffusion-limited kinetics. It overlaps with the rare-binding description of Cas9 activity when target search is slow ($k_{\text{PAM}} \ll k_{\text{unb}}$) and target recognition is not immediate ($k_{\text{R},i} \ll k_{\text{unb}}$). For abundant RNP (**Equation 2.3**), the above expression for $k_{\text{clv},i}$ may be rewritten in terms of RNP concentrations, yielding

$$k_{\text{clv},i} = \frac{[\text{RNP}]}{[\text{RNP}] + K_D} k_{\text{R},i} \quad \text{if } [\text{RNP}] \gg [\text{PAM}] + K_D \text{ and } [\text{RNP}] + K_D \gg k_{\text{R},i}/k_{\text{a}}^0. \tag{2.10}$$

We use this expression in **Chapter 3** to estimate the PAM dissociation constant K_D *in vitro* (using the biochemistry-aligned notation $k_{\text{on}}^0 = k_{\text{a}}^0$, $k_{\text{on}} = k_{\text{PAM}}$ and $k_{\text{off}} = k_{\text{unb}}$).

It should be noted that the name “PAM equilibrium” refers to the equilibrium between the unbound and PAM-bound state of a particular DNA site i, which is established at a rate $k_{\text{PAM}} + k_{\text{unb}}$. Our previous discussion of free-RNP concentrations assumed a different equilibrium, that between the collective of all RNPs and PAM sites, which is established at a rate $k_{\text{PAM}} + k_{\text{unb}} + k_{\text{bnd}}$ (**Suppl. Note 2.S3**). These two equilibria are typically established simultaneously, but they are not identical. For example, in an *in vivo* setting with scarce RNP ($k_{\text{PAM}} \ll k_{\text{unb}}$) and very dense DNA ($k_{\text{bnd}} \gg k_{\text{unb}}$), the RNPs may equilibrate faster than the timescale on which individual sites i attain a PAM equilibrium. Distinguishing these two equilibria is not merely semantic, but reflects the fundamental property of Cas9 that it reacts with diverse DNA sites at heterogeneous reaction rates $k_{\text{R},i}$.

2.5. Conclusion

With a minimal kinetic model of Cas9 activity, we have derived how the cleavage of a given DNA site follows from PAM binding, PAM unbinding, R-loop formation, RNP presence and PAM presence. Our expressions of the PAM-binding rate and the effective cleavage rate provide a general description of Cas9 cleavage while also delineating the kinetics limit of rare binding and PAM equilibrium that respectively apply in *in vivo* and *in vitro* environments. The theory of this chapter provides a fundament for the next chapters of this work, each of which is set in either environments, and it allows one to generalize the results beyond their original context.

2.6. Supplementary Notes

2.S1. Effective cleavage rate

We study $T_{\text{clv},i}$, the first passage time (FPT) to the cleaved state of site i by considering all possible state trajectories. In Laplace space, the distributions of PAM binding, PAM unbinding and R-loop formation are given by

$$\phi_{\text{PAM}}(s) = \frac{k_{\text{PAM}}}{k_{\text{PAM}} + s}, \tag{2.S1}$$

$$\phi_{\text{unb},i}(s) = \frac{k_{\text{unb}}}{k_{\text{unb}} + k_{\text{R},i} + s}, \tag{2.S2}$$

$$\phi_{R,i}(s) = \frac{k_{\text{R},i}}{k_{\text{unb}} + k_{\text{R},i} + s}. \tag{2.S3}$$

If Cas9, after initial association to the PAM, dissociates and re-associates exactly n times before successful cleavage, the convolution of these $2n + 2$ exponential processes gives the FPT distribution at the cleaved state $\psi^n_{\text{clv},i}$. In Laplace space, we multiply the individual processes to arrive at

$$\psi^n_{\text{clv},i}(s) = \phi_{\text{PAM}} \cdot \left(\phi_{\text{unb},i} \cdot \phi_{\text{PAM}}\right)^n \cdot \phi_{R,i}\,. \tag{2.S4}$$

We consider any number n of PAM unbinding and rebinding cycles and sum over n,

$$\psi_{\mathrm{clv},i}(s) = \sum_{n=0}^{\infty} \psi_{\mathrm{clv},i}^{n}(s) = \sum_{n=0}^{\infty} \phi_{\mathrm{PAM}} \cdot \left(\phi_{\mathrm{unb},i} \cdot \phi_{\mathrm{PAM}}\right)^{n} \cdot \phi_{R,i} = \frac{\phi_{\mathrm{PAM}} \cdot \phi_{R,i}}{1 - \phi_{\mathrm{PAM}} \cdot \phi_{\mathrm{unb},i}}, \quad (2.S5)$$

where we use the closed form of a geometric sum, $\sum_{n=0}^{\infty} x^n = 1/(1-x)$ if $|x| < 1$. In its current form, $\psi_{\mathrm{clv},i}(s) = P(s)/Q(s)$ is a polynomial fraction with a linear numerator $P(s)$ and quadratic denominator $Q(s)$. We obtain the distribution of $T_{\mathrm{clv},i}$ in temporal space $f_{\mathrm{clv},i}(t)$ by performing fractional decomposition before inverting the Laplace transform.

$$\psi_{\mathrm{clv,i}}(s) = \frac{k_{\mathrm{PAM}} \cdot k_{\mathrm{R},i}}{k_{1,i} - k_{2,i}} \left(\frac{1}{s + k_{2,i}} - \frac{1}{s + k_{1,i}} \right) \quad (2.S6)$$

$$f_{\mathrm{clv},i}(t) = \mathcal{L}^{-1}\left[\psi_{\mathrm{clv,i}}(s)\right](t) = \frac{k_{\mathrm{PAM}} \cdot k_{\mathrm{R},i}}{k_{1,i} - k_{2,i}} \left(e^{-k_{2,i}t} - e^{-k_{1,i}t}\right), \quad (2.S7)$$

where the rates $k_{1,i}$ and $k_{2,i}$ are the negative roots of $Q(s)$,

$$k_{1,i} = \frac{1}{2}\left(k_{\mathrm{PAM}} + k_{\mathrm{unb}} + k_{\mathrm{R},i} + \sqrt{(k_{\mathrm{PAM}} + k_{\mathrm{unb}} + k_{\mathrm{R},i})^2 - 4\,k_{\mathrm{PAM}} k_{\mathrm{R},i}}\right), \quad (2.S8)$$

$$k_{2,i} = \frac{1}{2}\left(k_{\mathrm{PAM}} + k_{\mathrm{unb}} + k_{\mathrm{R},i} - \sqrt{(k_{\mathrm{PAM}} + k_{\mathrm{unb}} + k_{\mathrm{R},i})^2 - 4\,k_{\mathrm{PAM}} k_{\mathrm{R},i}}\right). \quad (2.S9)$$

The above PDF corresponds to a hypoexponential distribution, $T_{\mathrm{clv},i} \sim \mathrm{Hypo}(k_{1,i}, k_{2,i})$, describing the convolution of two exponential processes with rates $k_{1,i}$ and $k_{2,i}$. The distribution has a mean

$$\langle T_{\mathrm{clv},i} \rangle = \frac{1}{k_{1,i}} + \frac{1}{k_{2,i}} = \frac{k_{\mathrm{PAM}} + k_{\mathrm{unb}} + k_{\mathrm{R},i}}{k_{\mathrm{PAM}} \cdot k_{\mathrm{R},i}}. \quad (2.S10)$$

When $k_{\mathrm{PAM}} \approx k_{\mathrm{R},i}$ while also $k_{\mathrm{PAM}} \gtrsim k_{\mathrm{unb}}$ and $k_{\mathrm{R},i} \gtrsim k_{\mathrm{unb}}$, the rates $k_{1,i}$ and $k_{2,i}$ are of similar magnitude, so one needs consider the full hypoexponential distribution (**Figure 2.3**). Otherwise, there is a separation of timescales $k_{1,i} \gg k_{2,i}$, such that at times $t \gg 1/k_{1,i}$ the distribution $f_{\mathrm{clv},i}(t)$ reduces to a simple exponential distribution $T_{\mathrm{clv},i} \sim \mathrm{Exp}(k_{2,i})$. In this case, the one relevant rate $k_{2,i}$ corresponds to the effective cleavage rate $k_{\mathrm{clv},i}$. From **Equation 2.S10** we obtain

$$k_{\mathrm{clv},i} = k_{2,i} \approx 1/\langle T_{\mathrm{clv},i} \rangle = \frac{k_{\mathrm{PAM}} \cdot k_{\mathrm{R},i}}{k_{\mathrm{PAM}} + k_{\mathrm{unb}} + k_{\mathrm{R},i}}. \quad (2.S11)$$

2.S2. RNP equilibrium concentration

We derive the free-RNP concentration under the assumption of an equilibrium between RNP in solution and PAM-bound RNP. RNP-PAM binding is described by the reaction

$$\mathrm{RNP}_{\mathrm{free}} + \mathrm{PAM}_{\mathrm{free}} \overset{K_D}{\longleftrightarrow} \mathrm{RNP{:}PAM}. \quad (2.S12)$$

The equilibrium concentrations can be expressed in terms of the equilibrium constant,

$$K_D = \frac{k_{\text{unb}}}{k_{\text{a}}^0} = \frac{[\text{RNP}_{\text{free}}]_{\text{eq}} \cdot [\text{PAM}_{\text{free}}]_{\text{eq}}}{[\text{RNP:PAM}]_{\text{eq}}}. \tag{2.S13}$$

For total concentrations $[\text{RNP}] = [\text{RNP}_{\text{free}}] + [\text{RNP:PAM}]$ and $[\text{PAM}] = [\text{PAM}_{\text{free}}] + [\text{RNP:PAM}]$, conservation of mass implies that

$$[\text{RNP:PAM}] = [\text{RNP}] - [\text{RNP}_{\text{free}}] \tag{2.S14}$$

$$[\text{PAM}_{\text{free}}] = [\text{PAM}] - [\text{RNP:PAM}] \tag{2.S15}$$

$$= [\text{PAM}] - ([\text{RNP}] - [\text{RNP}_{\text{free}}]). \tag{2.S16}$$

Substituting the above expressions in the equilibrium equation creates a quadratic equation,

$$0 = ([\text{RNP}_{\text{free}}]_{\text{eq}})^2 + ([\text{PAM}] - [\text{RNP}] + K_D) \cdot [\text{RNP}_{\text{free}}]_{\text{eq}} - [\text{RNP}]K_D, \tag{2.S17}$$

$$[\text{RNP}_{\text{free}}]_{\text{eq}} = \frac{1}{2}\left([\text{RNP}] - [\text{PAM}] - K_D + \sqrt{([\text{RNP}] - [\text{PAM}] - K_D)^2 + 4[\text{RNP}]K_D}\right) \tag{2.S18}$$

$$= \frac{1}{2}\left([\text{RNP}] - [\text{PAM}] - K_D + \sqrt{([\text{RNP}] + [\text{PAM}] + K_D)^2 - 4[\text{RNP}][\text{PAM}]}\right), \tag{2.S19}$$

where we have selected the larger solution of the quadratic formula as it has the appropriate range $[\text{RNP}_{\text{free}}]_{\text{eq}} \in (0, [\text{RNP}])$ to be the physical solution.

The corresponding equilibrium free-PAM concentration is

$$[\text{PAM}_{\text{free}}]_{\text{eq}} = \frac{1}{2}\left([\text{PAM}] - [\text{RNP}] - K_D + \sqrt{([\text{RNP}] + [\text{PAM}] + K_D)^2 - 4[\text{RNP}][\text{PAM}]}\right). \tag{2.S20}$$

2.S3. RNP equilibration rate

To derive a PAM-binding rate k_{PAM}, we have assumed a local equilibrium between free and PAM-bound RNPs. Here, we investigate if indeed such a local equilibrium is established faster than the rates $k_{\text{clv},i}$ at which targets get cleaved.

We first determine the full dynamics of RNP-PAM binding under the assumption that cleavage can be neglected. The binding reaction is described by

$$\frac{d}{dt}\begin{pmatrix}[\text{RNP}_{\text{free}}]\\ [\text{PAM}_{\text{free}}]\\ [\text{RNP:PAM}]\end{pmatrix} = \begin{pmatrix}-k_{\text{a}}^0[\text{RNP}_{\text{free}}][\text{PAM}_{\text{free}}] + k_{\text{unb}}[\text{RNP:PAM}]\\ -k_{\text{a}}^0[\text{RNP}_{\text{free}}][\text{PAM}_{\text{free}}] + k_{\text{unb}}[\text{RNP:PAM}]\\ +k_{\text{a}}^0[\text{RNP}_{\text{free}}][\text{PAM}_{\text{free}}] - k_{\text{unb}}[\text{RNP:PAM}]\end{pmatrix}. \tag{2.S21}$$

The above can be written as a single ODE for $y(t) = [\text{RNP:PAM}]$,

$$\frac{dy}{dt} = k_{\text{a}}^0 (y - y_1)(y - y_2), \tag{2.S22}$$

where the fixed points y_1 and y_2 correspond to the solutions to the quadratic equation formulated in **Equation 2.S17**,

$$\begin{aligned} y_1 &= \frac{1}{2}\left([\text{RNP}] + [\text{PAM}] + K_D + \sqrt{([\text{RNP}] + [\text{PAM}] + K_D)^2 - 4[\text{RNP}][\text{PAM}]}\right), \\ y_2 &= \frac{1}{2}\left([\text{RNP}] + [\text{PAM}] + K_D - \sqrt{([\text{RNP}] + [\text{PAM}] + K_D)^2 - 4[\text{RNP}][\text{PAM}]}\right). \end{aligned} \tag{2.S23}$$

The ODE is solved by separation of variables,

$$\int \frac{dy}{k_{\text{a}}^0 (y - y_1)(y - y_2)} = \int dt, \tag{2.S24}$$

$$\frac{1}{k_{\text{a}}^0 (y_1 - y_2)} \left(\int \frac{dy}{y - y_1} - \int \frac{dy}{y - y_2} \right) = t + \text{constant}, \tag{2.S25}$$

$$\frac{1}{k_{\text{a}}^0 (y_1 - y_2)} \log\left(\frac{y - y_2}{y - y_1}\right) = t + \text{constant}, \tag{2.S26}$$

$$y = y_2 + (y - y_1)\exp\left(-k_{\text{a}}^0 (y_1 - y_2)t\right) \times \text{constant}. \tag{2.S27}$$

Solving the integration constant for $y(0) = 0$ results in

$$y(t) = [\text{RNP}] - ([\text{RNP}] - [\text{RNP}_{\text{free}}]_{\text{eq}}) \left(1 - \frac{k_{\text{eq}} e^{-k_{\text{eq}} t}}{k_{\text{eq}} + k_{\text{a}}^0 ([\text{RNP}] - [\text{RNP}_{\text{free}}]_{\text{eq}})(1 - e^{-k_{\text{eq}} t})}\right). \tag{2.S28}$$

with a characteristic rate

$$\begin{aligned} k_{\text{eq}} &= k_{\text{a}}^0 \cdot \sqrt{([\text{RNP}] + [\text{PAM}] + K_D)^2 - 4[\text{RNP}][\text{PAM}]} && \text{(2.S29)} \\ &= k_{\text{a}}^0 \cdot \left([\text{RNP}_{\text{free}}]_{\text{eq}} + [\text{PAM}_{\text{free}}]_{\text{eq}} + K_D\right) && \text{(2.S30)} \\ &= k_{\text{PAM}} + k_{\text{bnd}} + k_{\text{unb}}. && \text{(2.S31)} \end{aligned}$$

Here we have used the single-RNP binding rate $k_{\text{bnd}} = k_{\text{a}}^0 \cdot [\text{PAM}_{\text{free}}]_{\text{eq}}$ that describes how fast a single RNP binds any PAM site. The RNP-equilibration rate k_{eq} is larger than the rate k_{PAM} at which individual PAMs get bound, and so it is larger than the rate at which DNA sites get cleaved,

$$k_{\text{eq}} > k_{\text{PAM}} > \frac{k_{\text{R},i}}{k_{\text{PAM}} + k_{\text{unb}} + k_{\text{R},i}} k_{\text{PAM}} = k_{\text{clv},i}. \tag{2.S32}$$

In many settings, one can even safely assert that $k_{\text{eq}} \gg k_{\text{clv},i}$. For instance, PAM

abundance $[\text{PAM}] \gg [\text{RNP}]$ implies $[\text{PAM}_{\text{free}}]_{\text{eq}} \gg [\text{RNP}_{\text{free}}]_{\text{eq}}$ such that $k_{\text{bnd}} \gg k_{\text{PAM}}$ and accordingly $k_{\text{eq}} \gg k_{\text{PAM}}$. At the other extreme, we get $k_{\text{eq}} \approx k_{\text{clv},i}$ when RNP is abundant ($[\text{RNP}] \gg [\text{PAM}] + K_D$ such that $k_{\text{eq}} \approx k_{\text{PAM}}$) and R-loop formation is fast ($k_{\text{R},i} \gg k_{\text{PAM}} + k_{\text{unb}}$ such that $k_{\text{PAM}} \approx k_{\text{clv},i}$). Even in this limit, our expression for k_{PAM} is valid: due to the RNP abundance, the free RNP concentration is always effectively equal to the total RNP concentration (**Equation 2.S35**) regardless of whether a local equilibrium is achieved.

In conclusion, our definition of k_{PAM} is consistent with our approach of deriving RNP availability from an assumed local RNP-PAM equilibrium.

2.S4. Equilibrium concentrations for abundant and scarce RNP and PAM

The equilibrium free-RNP concentration $[\text{RNP}_{\text{free}}]_{\text{eq}}$ simplifies in the limits of either RNP or PAM abundance. Rewriting **Equation 2.S19** in terms of the dimensionless variables $q = [\text{RNP}_{\text{free}}]_{\text{eq}}/[\text{RNP}]$, $x = [\text{PAM}]/[\text{RNP}]$ and $z = (K_D + [\text{PAM}])/[\text{RNP}]$, we obtain

$$q = \frac{1}{2}\left(1 - z + \sqrt{(1+z)^2 - 4x}\right). \tag{2.S33}$$

This formula can be simplified for small and large z by taking into account that $x < z$. This yields

$$q = \begin{cases} 1 & \text{if } z \ll 1, \\ x/z & \text{if } z \gg 1. \end{cases} \tag{2.S34}$$

Re-expressing this in terms of the original parameters, we find

$$[\text{RNP}_{\text{free}}]_{\text{eq}} = \begin{cases} [\text{RNP}] & \text{if } [\text{RNP}] \gg [\text{PAM}] + K_D, \quad (2.S35) \\ \dfrac{K_D}{[\text{PAM}] + K_D} \cdot [\text{RNP}] & \text{if } [\text{RNP}] \ll [\text{PAM}] + K_D. \quad (2.S36) \end{cases}$$

As RNP-PAM binding is symmetric, the same derivation applies to $[\text{PAM}_{\text{free}}]_{\text{eq}}$, such that **Equation 2.S20** simplifies to

$$[\text{PAM}_{\text{free}}]_{\text{eq}} = \begin{cases} [\text{PAM}] & \text{if } [\text{PAM}] \gg [\text{RNP}] + K_D, \quad (2.S37) \\ \dfrac{K_D}{[\text{RNP}] + K_D} \cdot [\text{PAM}] & \text{if } [\text{PAM}] \ll [\text{RNP}] + K_D. \quad (2.S38) \end{cases}$$

References

1. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology* **17,** 148. doi:10.1186/s13059-016-1012-2 (2016).
2. Moreb, E. A. & Lynch, M. D. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nature Communications* **12.** Number: 1, 5034. doi:10.1038/s41467-021-25339-3 (2021).

3. Cameron, P. et al. Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nature Methods* **14,** 600–606. doi:10.1038/nmeth.4284 (2017).

4. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences* **114,** 5461–5466. doi:10.1073/pnas.1700557114 (2017).

5. Boyle, E. A. et al. Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Science Advances* **7.** doi:10.1126/sciadv.abe5496 (2021).

6. Jones, D. L. et al. Kinetics of dCas9 target search in Escherichia coli. *Science* **357,** 1420–1424. doi:10.1126/science.aah7084 (2017).

7. Jones, S. K. et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nature Biotechnology* **39.** ISBN: 4158702006465, 84–93. doi:10.1038/s41587-020-0646-5 (2021).

8. Lu, Q. & Pigolotti, S. Target search in the CRISPR/Cas9 system: facilitated diffusion with target cues. *New Journal of Physics* **26,** 043007. doi:10.1088/1367-2630/ad3841 (2024).

9. Globyte, V., Lee, S. H., Bae, T., Kim, J.-S. & Joo, C. CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. *The EMBO Journal* **38.** doi:10.15252/embj.201899466 (2019).

10. Lu, Q., Bhat, D., Stepanenko, D. & Pigolotti, S. Search and Localization Dynamics of the CRISPR-Cas9 System. *Physical Review Letters* **127,** 208102. doi:10.1103/PhysRevLett.127.208102 (2021).

11. Martens, K. J. A. et al. Visualisation of dCas9 target search in vivo using an open-microscopy framework. *Nature Communications* **10.** Number: 1, 3552. doi:10.1038/s41467-019-11514-0 (2019).

3

CRISPRzip explains and predicts sequence- and context-dependent Cas9 activity

Hidde S. Offerhaus, Ieva Jaskovikaitė, Stephen K. Jones, Martin Depken

Accurate prediction of CRISPR-based gene editing remains challenging since existing models often fail to generalize across experimental and cellular contexts. We introduce CRISPRzip, a mechanistic kinetic model that quantitatively links nucleotide sequence and environmental conditions to target interrogation. The model describes R-loop formation as movement through a sequence-dependent free-energy landscape, combining nearest-neighbor nucleic-acid energetics with protein specific contributions inferred from high-throughput binding and cleavage kinetics. Applying CRISPRzip to SpCas9, we predict the activity across diverse DNA targets and guide RNAs, and validate with independent single-molecule FRET and torque spectroscopy experiments. By explicitly incorporating Cas9 concentration and DNA superhelicity, the framework predicts heterogeneous, context-dependent editing outcomes and rationalizes how physical constraints modulate cleavage dynamics. Our results provide a transferable, physics-based foundation that unifies mechanistic insight and predictive power, enabling robust characterization of CRISPR effectors and prediction of their activity across environmental contexts.

This chapter is under preparation for submission. A preprint is available on bioRxiv.

3.1. Introduction

RNA-guided nucleases (RGNs) from CRISPR immune systems are quintessential tools for genome editing [1], with applications ranging from GMO crops [2] to therapeutics [3–5]. Their specific activity is programmed by binding a guide RNA (gRNA) that complements the intended DNA target [6]. When the effector protein recognizes its target, it creates a double-stranded break (DSB) that can cause various edits as the cell repairs the DNA [7–9].

Not all gRNAs produce successful edits, either because they lack activity on the intended target or because they cause undesired edits on off-target sites [10–14]. These variable outcomes arise from a complex interplay between intrinsic factors such as the enzyme, gRNA, and target, and extrinsic factors such as delivery efficacy, cell-cycle and genomic state, and DNA repair pathways. Numerous models exist that predict the on- and off-target activity based on gRNA sequence, with the number of deep-learning models growing particularly fast [15–19]. These models are trained in a specific cellular context, and rapidly lose predictive power at other conditions [10, 16, 20]. This limited generalizability is a serious weakness given the diversity of CRISPR applications and of relevant experimental and therapeutic conditions.

Eslami-Mossallam et al. (2022) [21] introduced a kinetic model that makes quantitative activity predictions when trained on *Streptococcus pyogenes* Cas9 (SpCas9) data. It describes Cas9 dynamics in terms of the free-energy landscapes that govern target interrogation. The underlying physical parameters can be inferred from kinetic data of Cas9 binding and cleaving a large library of DNA targets. While this model deliveres considerable predictive power on the basis of mismatch positions alone, its insensitivity to nucleotide sequence renders it unable to capture the full variation in activity between DNA targets and RNA guides.

Here we present CRISPRzip, a general kinetic RGN framework that we apply to SpCas9, which integrates gRNA- and DNA-sequence dependence and captures variable nuclease concentration and DNA superhelicity.

CRISPRzip incorporates the free-energy cost of the strand-replacement reaction by which Cas9 establishes an R-loop between the gRNA and the target DNA. Previous biophysical models correlated the stability of a complete R-loop with Cas9 activity [22–25]. However, the R-loop reaction need not equilibrate before cleavage [26, 27], so the full kinetic pathway should be accounted for [28]. In doing so, we capture Cas9 activity across DNA and gRNA sequences, and identify anomalous DNA targets on which Cas9:gRNA may form sequence-specific interactions beyond the R-loop hybrid.

Our predicted landscapes explain dynamics across experimental modalities, which allows us to validate CRISPRzip by comparison to Förster resonance energy transfer (FRET) assays and single-molecule torque spectroscopy measurements. The model captures the effects of Cas9:gRNA concentration and DNA superhelicity in physically meaningful parameters, and predicts and rationalizes a heterogeneous response to both these extrinsic conditions. CRISPRzip's sequence sensitivity and ability to adapt to environments showcase its capacity to assist in gRNA design and advance CRISPR applications.

3.2. Results

3.2.1. Model definition

CRISPRzip represents the interrogation of a DNA target by the Cas9:gRNA complex in 22 partial R-loop states (**Figure 3.1A**). The complex initially binds to DNA by specifically interacting with the protospacer-adjacent motif (PAM) 5'-NGG-3' next to a target (the "PAM" state, with $b = 0$ hybrid base pairs) [29]. Next, Cas9 catalyzes the formation of an R-loop between its gRNA and the target DNA. During this process, the dsDNA separates into a displaced non-target strand (NTS) [30–32] and a target strand (TS) that hybridizes with the gRNA. The R-loop forms in a zipper-like manner by stepwise and reversible strand replacement, starting directly upstream of the PAM. After the first dsDNA base pair is broken to form one hybrid base pair (hybrid size $b = 1$), the hybrid can either be reversed or extended. If extended, the process repeats and a longer hybrid ($b = 2, 3, \ldots, 20$) can be formed [32–34]. We model R-loop extension from state $b - 1$ to b with a single forward rate k_{f} (**Table 3.S1**), and take R-loop recession from b to $b - 1$ to occur at a rate $k_{\mathsf{f}} \exp(-\Delta U_b / k_B T)$, where $\Delta U_b = U_b - U_{b-1}$ is the free-energy difference between the states. As the R-loop is completed, Cas9 undergoes a structural reconfiguration that exposes the active sites cleaving the TS and NTS [35, 36]. Experimental work shows that cleavage occurs with a rate of at least $k_{\mathsf{cat}} = 4.0\,\mathrm{s}^{-1}$ [37, 38]. We simulate it as an irreversible transition from the fully opened R-loop state ($b = 20$) to the cleaved state, and adopt the above k_{cat} value as a model constant (**Table 3.S1**). With the above parametrization, we can simulate R-loop kinetics at base pair resolution once we determine the unknown values of the free-energy landscape $U_b = (U_1, \ldots, U_{20})$ and the R-loop extension rate k_{f} (**Methods 3.4**).

Our model assumes that a free-energy landscape U_b can be split into the sequence-dependent stability of an isolated R-loop in solution (the R-loop cost U_b^{R}), and a sequence-independent contribution due to interactions with Cas9 (the protein contribution U_b^{P}, **Figure 3.1B**),

$$U_b = U_b^{\mathsf{R}} + U_b^{\mathsf{P}}. \tag{3.1}$$

The R-loop cost U_b^{R} of forming an R-loop intermediate of length b is estimated based on nearest-neighbor parameters determined from nucleic-acid melting experiments (**Figure 3.S1**, **Methods 3.4**) [39–43]. We follow the procedure from Alkan et al. (2018) [23] to estimate the stability of mismatching hybrid base pairs at standard conditions. The R-loop cost at lower salt levels [44, 45] and variable temperature [39] can be derived by uniformly scaling the standard-condition parameters by a context-dependent factor α.

The protein contribution U_b^{P} represents the total of sequence-independent interactions between Cas9 and the R-loop, including interactions with the DNA backbone and reorganizations of the protein structure that are coupled to R-loop dynamics. With a perfect (gRNA-matching) target, the protein contribution takes on the value $U_b^{\mathsf{P,pt}}$. A mismatch at position m in the R-loop adds a protein mismatch penalty Q_m^{P} to $U_b^{\mathsf{P,pt}}$ at all states $b \geq m$ (**Figure 3.1B**, **Methods 3.4**). Unlike the R-loop cost, the values of $U_b^{\mathsf{P,pt}}$ and Q_b^{P} are unknown and must be inferred from experimental data (**Table 3.S1**).

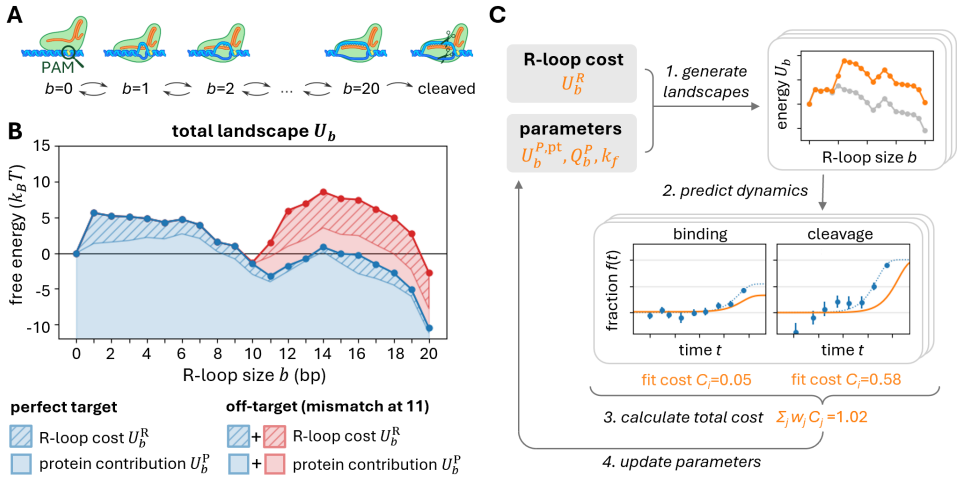


Figure 3.1: CRISPRzip model definition and training. **A** A 22-state kinetic model describes how Cas9 progressively forms an R-loop of $b = 0, \ldots, 20$ hybrid base pairs, followed by DNA cleavage. **B** Illustration of free-energy landscape contributions by different targets. Each landscape is split up into a known R-loop cost and an unknown protein contribution, according to $U_b = U_b^{\mathrm{R}} + U_b^{\mathrm{P}}$, with $U_b^{\mathrm{P}} = U_b^{\mathrm{P,pt}}$ for the perfect-target landscape (blue). The landscape of the off-target (blue + red) with a mismatch at position 11 has a higher R-loop cost compared to the perfect target (red, hatched), as well as a protein mismatch penalty Q_{11}^P that is included in the protein contribution U_b^{P} at all positions $b \geq 11$ (red). **C** A graphic overview of the training procedure for CRISPRzip's parameters. 1. Landscapes are created for all targets in the training set on the basis of the current set of parameters. 2. From the predicted free energy landscapes, binding and cleavage dynamics are simulated. 3. Each binding and cleavage curve is scored on its fit quality, and all scores are weighed and added up to a total cost. 4. On the basis of the total cost, the parameter set can be accepted or rejected according to simulated annealing. Steps 1–4 are repeated with newly updated parameter sets until convergence.

3.2.2. Obtaining training data with high-throughput kinetic assays

We measured how DNA targets are cleaved by Cas9 and bound by catalytically dead Cas9 (dCas9). To observe gRNA sequence effects, we performed high-throughput experiments on two distinct DNA libraries derived from DNA targets λ1 and λ2 of the λ-phage genome [46, 47], hereafter called target 1 and 2. Both libraries included the perfect target and all 1770 targets with one or two point mutations relative to the perfect target (**Table 3.S1**). To measure cleavage activity, we incubated each DNA library with Cas9:gRNA ribonucleoprotein (RNP) for durations from 10 seconds to 1000 minutes (logarithmically spaced) and quantified the cleaved fraction $f_{\mathrm{clv}}(t)$ of each target with next-generation sequencing (NGS; **Methods 3.4**, **Table 3.S1**). To measure binding by dCas9, we implemented a filtration assay [48] that separated PAM-bound targets from targets bound via sufficient R-loop formation. We incubated each DNA library with RNP at the same durations as above and quantified the bound fractions

$f_{\mathrm{bnd}}(t)$ by performing NGS on the filtrate (**Methods 3.4**). Our active RNP concentration (62.5 nM) was large enough to assume saturation, ensuring that effectively each library member was Cas9-bound. Experiments were carried out at room temperature and physiological salt levels (**Methods 3.4**), for which we estimate an R-loop scaling factor of $\alpha = 0.7 \pm 0.2$ (**Figure 3.S2**) based on DNA:RNA melting experiments at similar salt concentrations [45].

The CRISPRzip parameters $U_b^{\mathrm{P,pt}}$, Q_b^{P} and k_{f} can be inferred by minimizing the total least-squared errors between our predictions and the observed bound and cleaved fractions at all measured time points (**Figure 3.1C**, **Methods 3.4**). The training set consists of the perfect target, all single mismatch off-targets, and 90% of the double mismatch off-targets in a library. The model performance is evaluated with a test set consisting of the excluded 10% of all double mismatch off-targets, selected by stratified sampling.

3.2.3. Establishing a single model for variable-gRNA activity prediction

Before we establish our definitive parameter values, we validate if the R-loop cost improves predictions across multiple DNA and gRNA sequences. Next, based on our training results, we construct a single parameter set that generates smooth and interpretable free-energy landscapes while maintaining performance.

Calculated R-loop costs may explain the variations among targets given a single gRNA—for instance, between two off-targets with different mismatched bases at the same position. In order to assess CRISPRzip's capacity to capture mismatched-base effects, we set up a sequence-agnostic control model (**Figure 3.2A**). The control model's landscapes U_b do not contain an R-loop cost U_b^{R}, and its protein contribution U_b^{P} is trained to capture the average effects of mismatches based on their positions. This approach is identical to that presented in Eslami-Mossallam et al. (2022) [21]. The training results show that with no additional fit parameters, including R-loop costs reduces the total fit cost by 28% and 36% for library 1 and 2, respectively (**Figure 3.2B**). The obtained parameters (**Figure 3.S3**) perform comparatively in the test sets and the training sets (**Figure 3.S4A**), demonstrating their general validity.

To test how far our model performance extends beyond the experimental modality used for training, we compare our models to an independent dataset of Cas9 dynamics from Aguirre Rivera et al. (2024) [49]. Here, single-molecule FRET was combined with NGS to monitor R-loop completion by Cas9 on all 256 possible sequence variations of the four most PAM-distal DNA bases (**Figure 3.2C**). The gRNA used in these experiments is nearly identical to our gRNA 1, so we use our library 1-trained parameter set to calculate R-loop formation probabilities on their DNA target library (**Methods 3.4**). CRISPRzip predicts the experimental results considerably better than the control model (**Figure 3.2D**), showing that model improvements persist across experimental modalities. Moreover, our Pearson correlation coefficient $r = 0.81$ approaches the correlation between repeated FRET measurements ($r = 0.88$), showing that our model explains the majority of the variation observed beyond experimental noise.

Next, we test if including the R-loop cost also allows us to predict across gRNA

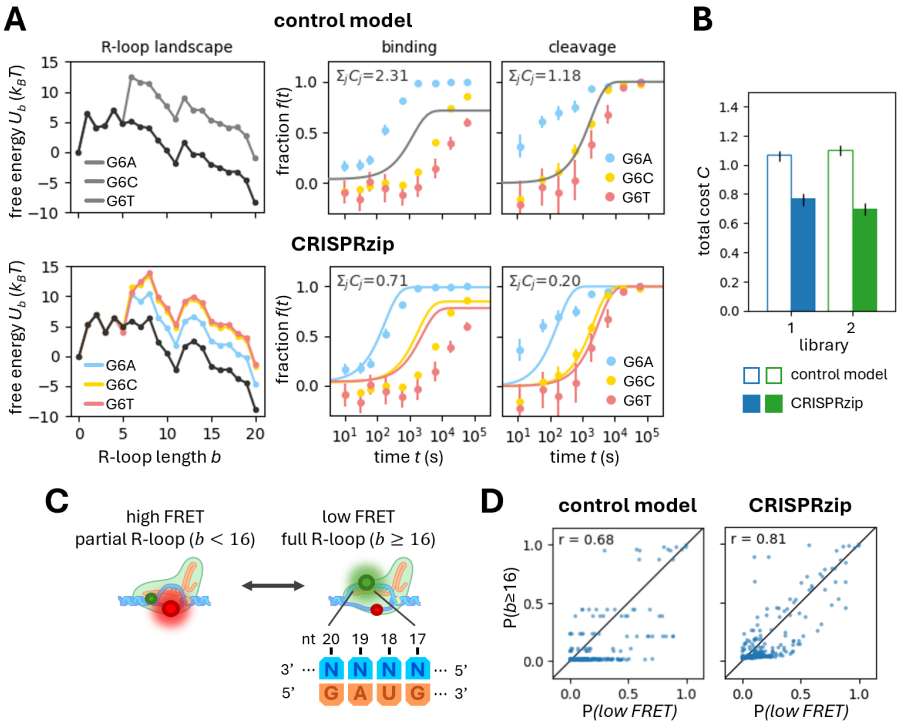


Figure 3.2: Establishing a single model for variable-gRNA activity prediction. A Example hybridization landscapes for the perfect target 2 (black) and three off-targets with mismatches at position 6, obtained by the reference model (top row) and with CRISPRzip (bottom row). **B** Results of training on the library 1 and 2 datasets with the control model (without R-loop cost) and CRISPRzip (with R-loop cost). Bars show average fit cost (train + test set) of the top 50% results from 200 optimization runs, error bars show 95% confidence interval. **C** Illustration of the high-throughput single-molecule FRET experiments performed by Aguirre Rivera et al. (2024) [49]. The target DNA is fluorescently labeled (Cy3 on the TS at $b = 6$ and Cy5 on the NTS at $b = 16$) such that the FRET signal drops upon R-loop completion ($b \geq 16$). Single-molecule FRET traces are obtained for all sequence variations of the four most PAM-distal DNA bases (blue) and a single gRNA (orange) close to our gRNA 1. **D** Correlations between $P(b \geq 16)$, the predicted probability to attain an R-loop of size $b \geq 16$, and $P(\textit{low FRET})$, the measured occupancy of the low-FRET state. Dots represent the 256 DNA library members, each with a different 4-nt PAM-distal target sequence. Pearson correlation coefficient r is reported.

sequences with a single model. We train CRISPRzip on libraries 1 and 2 simultaneously, and obtain parameter sets predicting perfect-target landscapes that vary with gRNA sequence (**Figure 3.3A**, **3.S3**). The performance after joint training approaches that after isolated training (**Figure 3.3B**) and is similar for the train and test targets (**Figure 3.S4B**). As a control, we apply the library 1 trained parameters to predict the library 2 data, and vice versa. This cross-prediction yields significantly worse-scoring fits (**Figure 3.3B**).

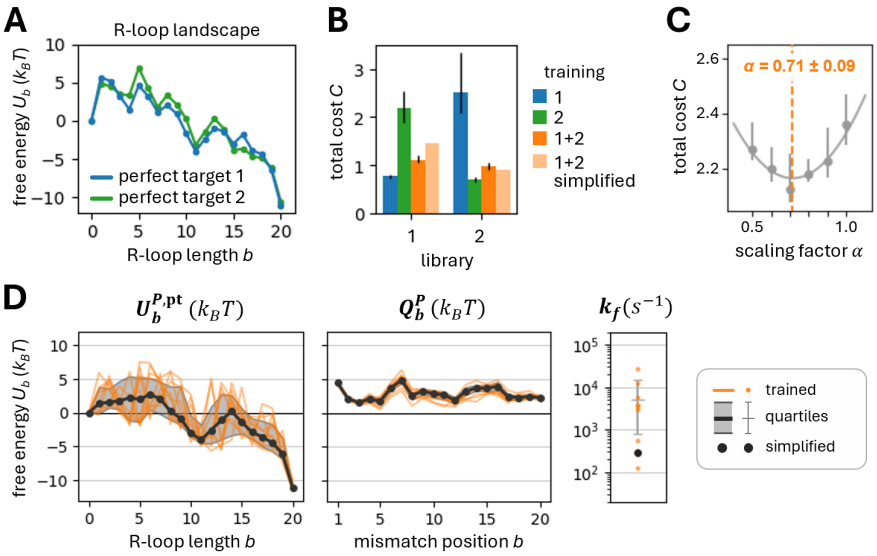


Figure 3.3: Establishing a single model for variable-gRNA activity prediction. **A** Joint training on library 1 and 2 predicts different R-loop landscapes for perfect target 1 (blue) and 2 (green). **B** Results of training on library 1 (blue), 2 (green), joint 1 and 2 (orange), and results of parameter simplification after joint training (light orange). Bars show average fit cost (train + test set) of the top 50% results from 200 optimization runs, error bars show 95% confidence interval. **C** Grid optimization of the R-loop scaling factor α. Parabolic interpolation approximates the optimum parameter value. **D** Simplification of parameters by first taking the median of $U_b^{\mathrm{P,pt}}$ (perfect-target protein contribution) and Q_b^{P} (protein mismatch penalties) from the the top-50% joint training results, and then re-optimizing the rate k_{f}. Orange lines and dots show 10 random individual training results, shaded region and error bar show the 25%, 50%, and 75% quartiles, black dots show the simplified parameter set.

Together, these observations indicate that training on multiple gRNAs increases the general applicability of the model without losing its predictive power for specific gRNAs.

Where we previously estimated the value of the R-loop scaling factor α on the basis of physiological-salt DNA:RNA melting experiments (**Figure 3.S2**), we now co-optimize it together with the jointly-trained model parameters. We perform grid optimization at $\alpha \in [0.5, 1.0]$ and estimate the optimum parameter value by parabolic interpolation (**Methods 3.4**). The optimal R-loop scaling factor $\alpha = 0.71 \pm 0.09$ (**Figure 3.3C**) closely agrees with our previous estimate, supporting the validity of the above parameter sets.

Most joint trainings consistently reproduce local energy minima in the perfect-target protein contribution $U_b^{\mathrm{P,pt}}$ at position 0, 11 and 20 (the “PAM”, “intermediate” and “open R-loop” states), but show considerable variation elsewhere (**Figure 3.3D**, **3.S3**). As our experiments cannot resolve individual R-loop steps, the measured effective transition dynamics between the metastable states can be realized by a variety of intervening barrier regions. Among the diversely shaped barrier regions in our training results, those

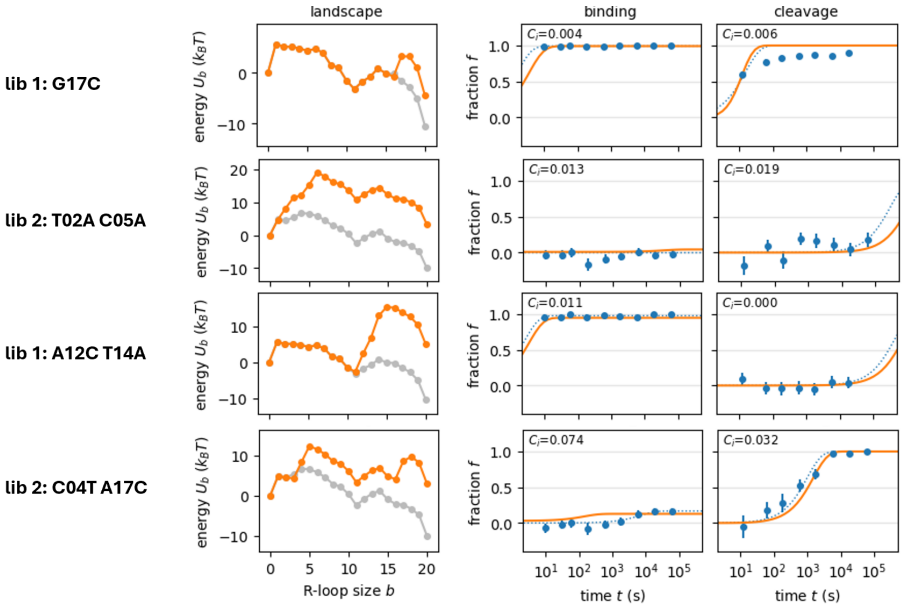


Figure 3.4: Free-energy landscapes explain variation in cleavage and binding dynamics. Predicted landscapes and kinetics (orange) of several off-targets, with perfect-target landscapes (gray) for reference. Experimental data (blue) with error bars indicating standard deviation and with exponential fit curve (dotted line). Fit cost C_i for a target i is the sum of squared errors of the CRISPRzip predictions compared to that of the exponential fit curve.

with high peaks are generally accompanied by large k_f values. To reveal common features across landscapes, we take the median of the model parameters $U_b^{\mathrm{P,pt}}$ and Q_b^{P} from the 50% best-scoring solutions (**Figure 3.3D**). As this median landscape is less peaked than the trained landscapes, we retrain the R-loop rate k_f (**Figure 3.3D**). The resulting simplified parameter almost completely maintains performance relative to the joint trainings (**Figure 3.3B**), only increasing the fit cost for binding double-mismatched library 1 targets (**Figure 3.S4B**). As we will argue below, our model's inability to capture the activity on these targets may arise from anomalous interactions beyond R-loop stability. All future analyses use this simplified parameter set because it makes consistent predictions while its smooth landscapes give visual insight into the effect of mismatches on R-loop dynamics.

3.2.4. Free-energy landscapes explain variation in cleavage and binding dynamics

The DNA targets in our libraries show a wide variety of binding and cleavage dynamics, much of which can be understood with CRISPRzip's predicted landscapes. The double-barrier structure of the protein contribution U_b^{P} generally carries over to the total R-loop landscape U_b. For many and diverse targets, the R-loop kinetics can be

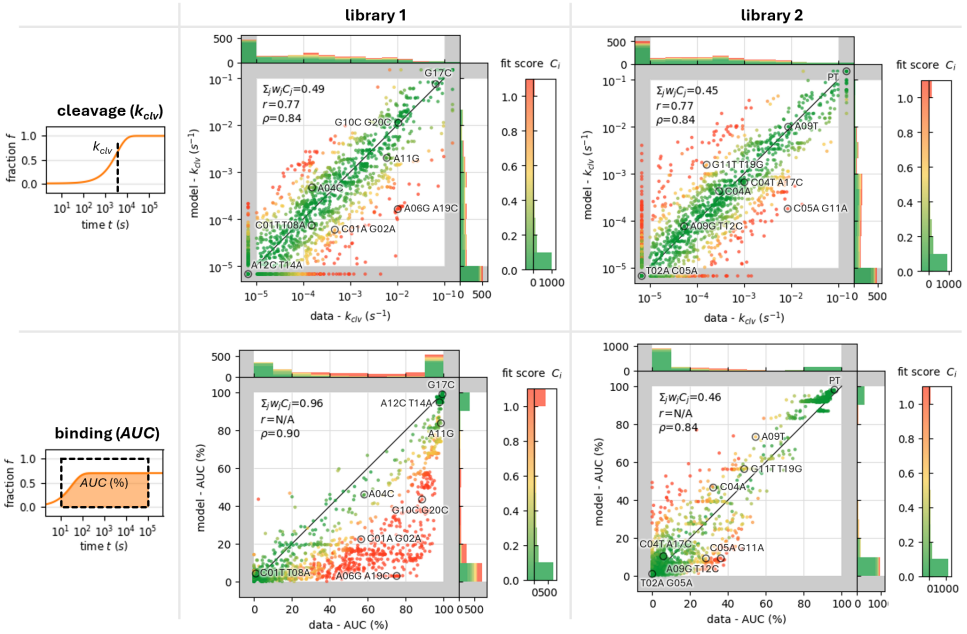


Figure 3.5: Free-energy landscapes explain variation in cleavage and binding dynamics. Correlations between the kinetic data and CRISPRzip predictions. Cleavage dynamics (top row) are expressed in terms of the effective cleavage rate k_{clv}, binding dynamics (bottom row) are summarized by the relative area under the curve (AUC; **Methods 3.4**). Values of k_{clv} outside the dynamic range $[10^{-5}, 10^{-1}]$ are collapsed to the gray border region. Colors indicate fit costs C_i of binding or cleavage fit curves. The total weighted sum of fit costs $\sum_j w_j C_j$ is reported for each of the four datasets. The Pearson correlation coefficient r is reported for the cleavage data (k_{clv} trimmed to the dynamic range), but not for the binding data because its near-bimodal AUC distributions violate the coefficient's requirement of approximate normality. Spearman's rank correlation coefficient ρ is reported for all data (with k_{clv} clipped to the dynamic range). Circles highlight the 16 selected targets for which example landscapes, binding curves and cleavage curves are shown (**Figure 3.4**, **3.S5**).

explained by simple landscape features: the height of both barriers and the stability of the intermediate and final (open R-loop) state (**Figure 3.4**).

We assess CRISPRzip's capacity to capture the experimental variation by visualizing its activity predictions on the full target libraries (**Figure 3.5**). The cleavage kinetics are summarized in terms of the effective cleavage rate k_{clv}, which is fit to the experimental and simulated data (**Methods 3.4**). The binding kinetics are summarized in terms of the area under the curve (AUC) of fit functions to the experimental and simulated data, which reflect both the rate of binding k_{bnd} and the final binding level ϕ_{bnd} (**Methods 3.4**). In the supplementary figures, a representative selection of targets is presented in closer detail to illustrate the relation between the R-loop landscapes and the predicted binding and cleavage kinetics (**Figure 3.S5**). The predictions generally correlate well to the kinetic data, reproducing cleavage and binding dynamics across a wide

range of timescales. Standard correlation metrics do not accurately represent model performance, as Pearson's r excludes the many k_{clv} values outside the dynamic range and is invalid for the bimodal AUC distributions, and Spearman's ρ is dominated by boundary clusters and is insensitive to the linear relation between model and data values. Instead, the weighted total fit cost $\Sigma_j w_j C_j$ provides the best indicator of quantitative agreement between predicted and experimental k_{clv} and AUC values.

For a group of 441 targets in library 1, making up 25% of the library, CRISPRzip consistently underestimates the binding level ϕ_{bnd} (**Figure 3.S6A,B**). 92% of these targets are characterized by one of 17 specific mutations at positions 2–10 (G02C/T, C03A/T, A04T, G05A/T, A06G, G07A, T08A/C, T08G, A09C/G, G10A/C/T), followed by any mismatch at position 11–20 (**Figure 3.S6C**). Among targets with this pattern, 80% is part of the underestimated-binding group. Interestingly, we correctly reproduce the cleavage dynamics of all anomalous targets, with the exception that targets with the mutation A06G are cleaved faster than predicted. We hypothesize that these mutations stabilize the intermediate state through protein-sequence interactions not captured by CRISPRzip.

3.2.5. CRISPRzip adapts to low Cas9 concentrations

Although our landscapes are trained with saturating Cas9 concentrations, we can generalize CRISPRzip to describe the dynamics at lower concentrations. We extend our 22-state model with an initial state representing an unbound DNA target. The kinetics of PAM-binding and -unbinding are described by the rates $k_{on} = k^0_{on} \cdot c$ and k_{off} respectively, with c denoting RNP concentration and k^0_{on} the molar PAM-binding rate. In order to determine the PAM dissociation constant $K_D = k_{off}/k^0_{on}$, we performed cleavage assays on three library-1 targets (perfect target (PT), G5A, C3T) at 7 RNP concentrations, ranging from 0.1 nM to 2 μM (**Methods 3.4**). We globally fit CRISPRzip predictions to the cleaved fractions and obtain a value of $K_D = 29.7 \pm 4.3$ nM (**Figure 3.6A**). These findings are implemented in CRISPRzip with the parameters values $k^0_{on} = 3.4 \cdot 10^{-2}$ nM^{-1} s^{-1} and $k_{off} = 1.0$ s^{-1} (**Table 3.S1**), in line with the PAM-bound Cas9 dwell times observed in *in vitro* single-molecule FRET experiments [34].

When we predict the effective cleavage rate k_{clv} for all single-mismatch targets in library E, we see that Cas9 concentrations do not affect targets equally (**Figure 3.6B**). For the targets with mismatches at positions 1–7 (including PT, G5A and C3T), cleavage follows Michealis-Menten kinetics (**Chapter 2**, **Equation 2.10**),

$$k_{clv}(c) = \frac{c}{K_D + c} \cdot k^{max}_{clv}. \tag{3.2}$$

Their first landscape barriers dominate the R-loop kinetics, hence c modulates their cleavage rate k_{clv} through the level of PAM-binding. Mismatches present at positions 12–16 create a rate-limiting PAM-distal barrier, where occupancy of the intermediate state, instead of the PAM state, defines the cleavage rates. Targets with single mismatches at positions 8–11 or 17–20 display intermediate concentration dependence as neither barrier dominates.

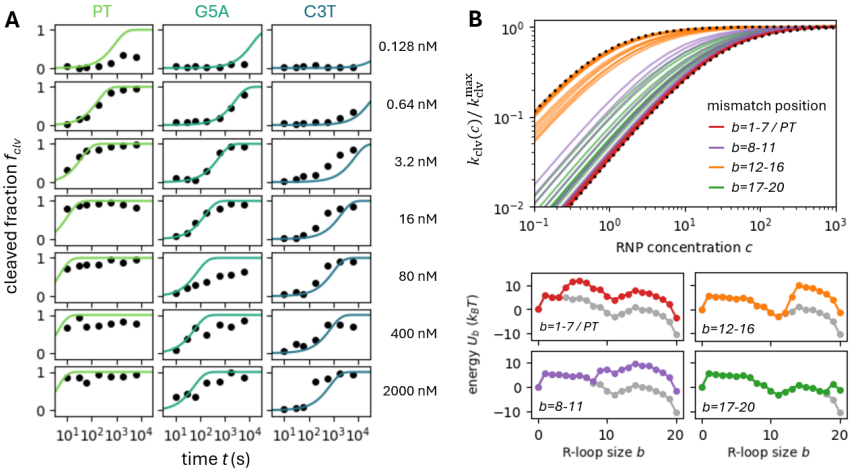


Figure 3.6: CRISPRzip adapts to low Cas9 concentrations. A A global fit to the cleavage of the perfect target (PT) and off-targets G5A and C3T from library 1 at various RNP concentrations c. Dots show data from filtration assay, lines show CRISPRzip predictions. **B** Simulated values of effective cleavage rate k_{clv} as a function of RNP concentration for the perfect target and all single-mismatch targets in library E. All k_{clv} values are shown relative to the maximum cleavage rate k_{clv}^{max} at saturating concentrations. Bottom dashed curve corresponds to **Equation 3.2**, top dashed curve shows the same relation for the effective dissociation constant $K_D^{eff} = k_{off}/k_{on}^0 \cdot (\sum_{b=10}^{13} \exp(-U_b/k_BT))^{-1} = 0.8\,\text{nM}$, indicating the maximum intermediate state stability. Example landscapes are shown for each mismatch region.

3.2.6. CRISPRzip adapts to DNA superhelicity

DNA superhelicity alters Cas9's capacity to cleave and bind targets [50]. Thanks to the physical basis of CRISPRzip, we can adapt it to capture the energetic effects of DNA superhelicity on R-loop formation (**Figure 3.7A**). To simulate the effects of a torque τ on a DNA target, the free-energy landscape U_b includes the work to extend the R-loop by one base pair, which requires helix unwinding over an angular distance θ (**Methods 3.4**). The torque may affect the R-loop extension rates too, so these are updated according to the work to reach the transition state between states b and $b+1$ at an angle φ.

In Ivanov et al. (2020) [47], magnetic tweezers experiments reveal the single-molecule kinetics of Cas9 on perfect target 1 at variable torque levels. The experiments indicate that R-loop formation requires full B-DNA unwinding, yielding $\theta = 34°$ (**Figure 3.S7**). Next, the study reports the effective rates between the PAM state "P" ($b = 0$), the intermediate state "I" ($b = 11$), and the open R-loop state "O" ($b = 20$) (**Figure 3.7B**). When we fit our model to the measurements at variable torque (**Methods 3.4**), we obtain a transition state angle of $\varphi = 50 \pm 5°$. The large value of φ compared to θ suggests that extending the R-loop requires temporary melting of the DNA also beyond the new hybrid base pair. As a consequence, the forward and backward rates between the R-loop states both slow down with positive torque and speed up with negative torque. With only one new fit parameter, CRISPRzip largely reproduces the data and captures

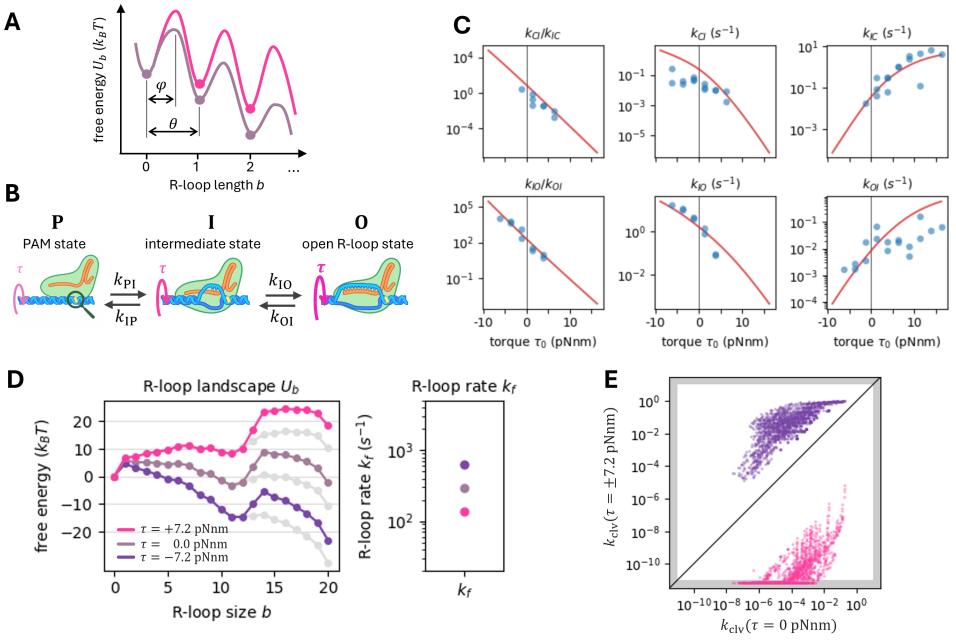


Figure 3.7: CRISPRzip adapts to DNA superhelicity. **A** Energetic effects of torque on the free-energy landscape U_b and its transition barriers. Lilac and pink lines illustrate torque-free and positive-torque landscapes, respectively. **B** Illustration of the observed metastable states and their effective transition rates. R-loop formation may be coupled to an increase in torque, depending on the DNA substrate. **C** Fitting the single-molecule twisting experiments of Ivanov et al. (2020) [47]. Experimental data is displayed as blue dots, the CRISPRzip fit is shown as a red line. The x-axis reports torque τ_0 on unopened DNA, which increases by up to 2.8 pN nm during R-loop formation. **D** Energy landscapes U_b and R-loop rate k_f for perfect target 1 (light gray) and its A13T off-target at constant torque $\tau = +7.2$ pN nm (pink), at zero torque (gray), and at $\tau = -7.2$ pN nm (purple), with the latter corresponding to bacterial genomes at superhelical density $\sigma = -0.05$. **E** Predictions of cleavage activity (k_{clv}) on all library-1 targets, comparing relaxed DNA ($\tau = 0$ pN nm) to DNA under negative ($\tau = -7.2$ pN nm, purple) and positive torque ($\tau = +7.2$ pN nm, pink).

the nonlinear dependence on torque (**Figure 3.7C**). This nonlinear effect arises from shifting the position of the highest barrier between the metastable states as torque increases (**Figure 3.7D**).

With these torsion parameters, we predict cleavage activity on the library-1 targets at an estimated bacterial genome torque of $\tau = -7.2$ pN nm (**Methods 3.4**) and at a positive torque of $\tau = +7.2$ pN nm. Our results show that torque alters cleavage rates k_{clv} heterogeneously (**Figure 3.7E**). For instance, negative torque accelerates the cleavage of targets with a factor 100–30,000, depending on the relative height of the two landscape barriers.

3.3. Discussion

CRISPRzip is a kinetic model that mechanistically captures the effects of nucleotide sequence and experimental context on RGN target recognition. By accounting for the progressive cost to form an R-loop, it quantitatively captures the variation across different DNA target and gRNA sequences when trained to SpCas9 binding and cleavage data. We validate the model across experimental modalities. CRISPRzip generates interpretable free-energy landscapes that illustrate target interrogation kinetics, and we isolate anomalous targets where Cas9 may interact specifically with the nucleic acid beyond R-loop base pairing. The environment heterogeneously affects Cas9's cleavage dynamics, as we demonstrate for both variable Cas9 concentration and variable DNA superhelicity. Adaptation to these extrinsic factors exemplify CRISPRzip's ability to model and predict activity in diverse applications.

A central advantage of CRISPRzip is that it does not only predict activities, but also returns protein-specific interaction-energy readouts that summarize how the protein shapes the R-loop formation pathway. In particular, the protein contribution provides an interpretable decomposition into on-target efficiency and off-target discrimination along R-loop formation (**Figure 3.3D**). These quantities can serve as context-independent selection criteria or engineering objectives, as variants may be benchmarked by how they increase mismatch penalties without compromising productive R-loop progression on targets.

Our capacity to capture sequence-specific effects relies on several factors. First, we need accurate nearest-neighbor parameters to calculate the true R-loop cost. Currently, we uniformly scale standard-condition parameters to estimate nucleic acid stability at physiological salt levels, but this simple scaling disregards the details of how base pair stabilities change with salt [51]. Optimal model performance would require that free energies of matching DNA:DNA and mismatching RNA:DNA duplexes would be determined at these conditions in nucleic-acid melting experiments. Alternatively, molecular dynamics could estimate these parameters [52, 53], and reveal how the presence of the Cas9 protein modifies the sequence-specific R-loop cost, as Cas9 bends DNA [35, 54, 55], holds the gRNA-TS hybrid in A-form [31], fixes the NTS backbone [35] and sterically constrains the nucleic acids.

Second, to obtain the protein contributions to the free energy landscapes, we should train on a set of gRNAs and DNA targets that more broadly capture the sequence space. While our current data sets comprise many DNA sequences, it would be informative to train on more than two gRNAs, especially if they have sequence features like GC-content that differs from our gRNAs 1 and 2 (**Figure 3.S8**). Previous work presented a method to quantify Cas9 binding and cleavage dynamics for many gRNAs in parallel [56], however we could not reproduce the reported gRNA 1 and 2 kinetics and thus did not train with it.

Third, we need experimental verification that sequence-specific R-loop costs translate to changes in the total free energy landscape of target interrogation. Single-molecule assays probe the stabilities of semi-stable landscape states for Cas9 and other Cas nucleases [34, 47, 57–59], with recent single-molecule work revealing the full energy landscape of R-loop formation by Cascade [60]. With these techniques, one can directly observe the influence of DNA and gRNA sequence on Cas9 kinetics, and hence provide

a direct test for an energetic model of sequence effects. For instance, these approaches could indicate if the anomalous binding patterns that we observed arise from non-local mismatch effects.

Although the R-loop stability is a major contributor to sequence dependence, the Cas9 protein can also specifically interact with the DNA and gRNA sequence when bases form contacts with the protein structure or when the R-loop geometry is distorted [55, 61]. These interactions may sometimes dominate the target interrogation dynamics but currently lie beyond CRISPRzip's scope. Still, CRISPRzip may be complemented with neural networks to recognize such sequence features and correct its predictions accordingly. Other off-target features that CRISPRzip does not account for include non-canonical PAMs [25, 62–64] and off-targets with missing or added base pairs in comparison to the perfect target [63]. These effects could be captured with a mechanistically motivated extension to the model. However, such an extension should balance parameter expansion with increased model performance and retained clarity.

With our explicit consideration of extrinsic factors and their physical effects, we contribute to a molecular understanding of the complex cellular processes that determine editing outcomes. For instance, Cas9 activity is expected to be strongly influenced by the local torsion state in a genome: transcription induces negative supercoiling [65], which correlates with increased Cas9 efficiency [50]. In **Chapter 4**, we present the target 1 and 2 libraries on supercoiled plasmids and quantify how Cas9 cleaves these. CRISPRzip simulates and rationalizes these results thanks to its torque dependence. Additionally, Cas9 delivery impacts editing outcomes due to the interplay between delivery strategies and cellular transport pathways [66]. By combining our mechanistic model with quantitative studies of cellular Cas9 delivery [67], we expect that editing outcomes can be explained in physical terms such as the nuclear RNP concentration profile as a function of time. Lastly, Moreb and Lynch (2020) [68] proposed that genetic context is a major determinant of activity, as many partially matching DNA sites may trap Cas9 and delay target search [10]. When applied over a whole genome, CRISPRzip can estimate targets' trapping capacities and predict their collective impact, again advancing insight into genome-scale phenomena from the bottom up.

In terms of activity prediction, we have previously demonstrated that universal mechanistic insight into the dynamics of Cas9 enables powerful and general off-target classification [21]. CRISPRzip also offers Cas9 activity prediction that can be applied to any genome. It is made publicly available as a graphical user interface [69], where users can provide gRNA and DNA sequences to inspect the predicted energy landscapes and binding and cleavage kinetics. We also provide a standalone Python package [70], which can form a biophysical foundation for more complex machine learning models. Recent computational work [71] demonstrated the promise of integrating kinetic models in a deep learning scheme to account for sequence-specific effects and cellular intricacies that are not feasible to capture mechanistically. Moreover, our model's applications could extend beyond wild-type SpCas9, to other Cas nucleases [72], engineered high-fidelity Cas9 variants [73], relaxed-PAM variants [74], and (d)Cas9 chimera like base and prime editors [1], provided sufficient data is available to train the protein contribution to the free-energy landscape. We believe that CRISPRzip provides a concrete solution for reliable gRNA design, and facilitates the development of safe and effective CRISPR applications.

3.4. Methods

3.4.1. Kinetic model

CRISPRzip derives the target interrogation kinetics from the free energy landscape of R-loop formation (**Figure 3.1A**). In total, there are 23 states $b \in \{-1, \dots, 21\}$ that each DNA target can occupy: the unbound state ($b = -1$), PAM-bound state ($b = 0$), the partial R-loop states ($b \in \{1, \dots, 20\}$) and the cleaved state ($b = 21$). At saturating Cas9 concentrations, the unbound state may be omitted from the system. Given the rates between the above states, we can solve the dynamics of R-loop formation. We parametrize these rates with an approach adopted from Eslami-Mossallam et al. (2022) [21]. The rates of PAM binding (from state -1 to 0), PAM unbinding (from 0 to -1) and strand cleavage (from 20 to 21) are defined directly as k_{on}, k_{off} and k_{cat}. The PAM-binding rate $k_{\text{on}} = k^0_{\text{on}} \cdot c$ follows from the molar PAM-binding rate k^0_{on} (units $\text{nM}^{-1}\,\text{s}^{-1}$) and the Cas9 concentration c (units nM). The rates of R-loop opening and closing between states $b \in \{0, \dots, 20\}$ are defined in terms of the free-energy landscape U_b,

$$U_b = U_b^{\text{R}} + U_b^{\text{P,pt}} + \sum_{m=0}^{b} Q_m^{\text{P}} \cdot \delta_m, \tag{3.3}$$

where U_b^{R} is the R-loop cost, $U_b^{\text{P,pt}}$ are the perfect-target protein contributions and Q_b^{P} are the protein mismatch penalties. The mismatch indicator δ_m takes on the value $\delta_m = 0$ in absence of a mismatch at position m, and $\delta_m = 1$ in presence of a mismatch, such that in the latter case the penalty Q_m^{P} is added to all states $b \geq m$. R-loop extension (from state b to $b+1$) takes place at a constant rate k_{f}. The R-loop recession rates (from b to $b-1$) follow from the free energy difference $\Delta U_b = U_b - U_{b-1}$. Together, these parameters yield a Markov model described by the forward and backward rates

$$k_{b\to(b+1)} = \begin{cases} k^0_{\text{on}} \cdot [\text{Cas9}] & \text{if } b = -1, \\ k_{\text{f}} & \text{if } b \in \{0, \dots 19\}, \\ k_{\text{cat}} & \text{if } b = 20, \end{cases} \tag{3.4}$$

$$k_{b\to(b-1)} = \begin{cases} k_{\text{off}} & \text{if } b = 0, \\ k_{\text{f}} \exp\left(\Delta U_b / k_B T\right) & \text{if } b \in \{1, \dots 20\}, \\ 0 & \text{if } b = 21. \end{cases} \tag{3.5}$$

These rates form the 23×23 transition rate matrix $\mathbf{K}$ of the system, defined as

$$K_{ij} = \begin{cases} k_{i\to j} & \text{if } i \neq j, \\ -\sum_m k_{i\to m} & \text{if } i = j. \end{cases} \tag{3.6}$$

The occupancy of all states $\mathbf{P}(t) = (P_{-1}(t), \dots, P_{21}(t))$ follows from the Master Equation $d\mathbf{P}/dt = \mathbf{K} \cdot \mathbf{P}$ with the initial condition that $\mathbf{P}(0) = (1, 0, \dots, 0)$. The Master

Equation is solved by

$$\begin{aligned}\mathbf{P}(t) &= \exp\left(\mathbf{K}t\right) \cdot \mathbf{P}(0) \\ &= \exp\left(\mathbf{U} \cdot \mathrm{diag}(\lambda_1 t, \ldots, \lambda_{23} t) \cdot \mathbf{U}^{-1}\right) \cdot \mathbf{P}(0) \\ &= \mathbf{U} \cdot \mathrm{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_{23} t}) \cdot \mathbf{U}^{-1} \cdot \mathbf{P}(0).\end{aligned} \tag{3.7}$$

The matrix exponential is sped up by the diagonalization of the transition rate matrix $\mathbf{K}$, writing its eigenvalues as $\{\lambda_1, \ldots, \lambda_{23}\}$ and its eigenvectors as the column vectors that make up $\mathbf{U}$. CRISPRzip's model definition and kinetic simulations are publicly available as a Python package [70].

3.4.2. R-loop cost

To calculate the R-loop cost, we follow the approach of Alkan et al. (2018) [23]. and sum over nearest-neighbor stability parameters for matching DNA:DNA base pairs [39], matching [40] and a few mismatching [41, 42] RNA:DNA base pairs. For the remaining mismatching RNA:DNA basepairs, the stabilities are estimated with an interpolation between DNA:DNA [39] and RNA:RNA [43] nearest-neighbor parameters. All nearest-neighbor contributions to the R-loop cost are taken at standard conditions (37 °C, 1 M NaCl), multiplied by the R-loop cost scaling factor α, and converted to units $k_B T$ (at room temperature $T = 22$ °C). Our calculations of intermediate R-loop costs take into account that there is a lag between DNA base stack opening and hybrid base stack formation (**Figure 3.S1**); to extend an R-loop from size $b-1$ to size b, the base stack between DNA bps $b-1$ and b are broken up and the base stack between hybrid bps $b-2$ and $b-1$ is formed. Similarly, to initiate R-loop formation ($b = 1$), two DNA base stacks are broken up with no hybrid base stack in return. Our R-loop cost calculations are available in the Python package [70].

3.4.3. Library preparation, cleavage and binding assays

The libraries 1 and 2 were designed as described in Hawkins et al. (2018) [75] and Jones et al. (2021) [63], and contain on-target sequences, non-target control sequences and thousands of off-target sequences with alternative PAMs or single and double mismatches, insertions, or deletions, when compared to the sgRNA. Each member has a unique barcode on each side, designed such that each cleavage product could be identified during sequencing. Each library was synthesized as pooled ssDNA oligonucleotides (GenScript), then amplified via PCR (Phusion Plus polymerase, Thermo Scientific; **Table 3.S4**, primers Pr6.29/Pr6.30, Azenta Life Sciences).

For NucleaSeq cleavage experiments, described in **Chapter 4** (**Methods 4.4**) unmodified sgRNAs (**Table 3.S3**) were purchased from Synthego or GenScript, and Cas9 and dCas9 from NEB. The sgRNA, (d)Cas9 and library DNAs were diluted in 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 100 μg mL^{-1} recombinant albumin, pH=7.9 (r3.1 buffer, NEB). RNP complex was formed by incubating Cas9 and sgRNA at 22 °C for 15 minutes, and then library DNA was added (sgRNA:Cas9:DNA of 187.5:62.5:6.25 nM, respectively) to initiate the cleavage experiment at 22 °C. The

reaction was sampled and quenched periodically (0, 0.2, 0.5, 1, 3, 10, 30, 100, 300 and 1000 min) in stop buffer (60 mM EDTA, 1 U Proteinase K, Thermo Scientific) for 30 minutes at 37 °C.

Binding affinity experiments combine elements of NucleaSeq library design and data processing with a previously described massively-parallel filter binding assay [48, 63]. The RNPs were formed and mixed with library DNA as in the cleavage experiments. The reaction was sampled at the same periodicity and stopped by passing through a 0.45 μM nitrocellulose filter (Cytiva) under vacuum.

DNA samples from cleavage and binding affinity experiments were prepared and sequenced by first appending time barcodes (indexes) with NGS prep (NEBNext, NEB), size-selecting for targets with magnetic beads (beads:DNA 0.9:1.0; AMPure XP, Beckman Coulter) and running on capillary electrophoresis (BioAnalyzer 2100, Agilent) to confirm size and purity of samples. The sample DNAs were sequenced along with 10% PhiX DNA (to ensure sequence diversity) at the EMBL GeneCore (Heidelberg, DE) on a NextSeq2000 using P2 100bp PE or P3 100bp PE chemistry.

3.4.4. Determination of binding and cleavage kinetics

Read counts are determined for each library member as previously described [63] (**Chapter 4**, **Methods 4.4**). In brief, each pair of reads from the sequencing reaction are merged, and then filtered by quality, agreement between member components, and size (to separate cut and uncut DNAs). The collated read counts $c_j(t)$ of library member j are expressed relative to its initial counts $c_j(0)$, and are normalized to changes in library composition over time (using non-target member counts, as they are not cut by Cas9). Hence, the fraction $f_j(t)$ of which library member j is bound or cleaved is calculated as

$$f_j(t) = 1 - \frac{c_j(t)/c_j(0)}{\langle c_0(t)/c_0(0) \rangle} \tag{3.8}$$

where $\langle c_0(t)/c_0(0) \rangle$ is the mean number of non-target counts relative to $t = 0$. All library members (including non-targets) with fewer than 50 initial counts are excluded from analysis. The errors in $f_j(t)$ are estimated by assuming Poisson distributed counts c_j with inferred standard deviation $\hat{\sigma} = \max\left(1, \sqrt{c_j}\right)$ and propagating the errors from all four count terms in **Equation 3.8**. Errors of $f_j(0)$ are zero by definition. When multiple library members j contain the same target sequence i, we calculate the target bound or cleaved fraction $f_i(t)$ and its error by taking an error-weighted average of the fractions $f_j(t)$ and their errors. Finally, we fit exponential curves to the binding and cleavage data,

$$y_{\text{bnd},i}(t) = \phi_{\text{bnd},i}\left(1 - \exp\left(-k_{\text{bnd},i} \cdot t\right)\right), \tag{3.9}$$

$$y_{\text{clv},i}(t) = 1 - \exp\left(-k_{\text{clv},i} \cdot t\right), \tag{3.10}$$

and express their fit quality in terms of residual sum of squares $\text{RSS}_i = \sum_t \left(y(t) - f_i(t)\right)^2$.

3.4.5. CRISPRzip training

We optimize CRISPRzip's parameters to fit the experimentally obtained binding and cleavage kinetics. The binding and cleavage data are split into a train and test set, with the test set comprising 10% of double-mismatch targets, obtained by stratified sampling of $\phi_{\text{bnd},i}$ and $k_{\text{bnd},i}$ (binding) or $k_{\text{clv},i}$ (cleavage). The 41 model parameters comprise 20 perfect-target protein contribution differences ($\Delta U_b^{\text{P,pt}} = U_b^{\text{P,pt}} - U_{b-1}^{\text{P,pt}}$, $b \in \{1, \dots, 20\}$), 20 protein mismatch penalties (Q_b^{P}, $b \in \{1, \dots, 20\}$) and the forward rate k_{f}. At each optimization step, an updated set of parameters predicts bound and cleaved fractions $\hat{f}_{\text{bnd},i}(t)$ and $\hat{f}_{\text{clv},i}(t)$ for each target i in the training set. The accuracy of the predictions is summarized in the cost function,

$$\begin{aligned} C = &\sum_i w_i \left(\sum_t \left(\hat{f}_{\text{bnd},i}(t) - f_{\text{bnd},i}(t) \right)^2 - \text{RSS}_{\text{bnd},i} \right) + \\ &\sum_i w_i \left(\sum_t \left(\hat{f}_{\text{clv},i}(t) - f_{\text{clv},i}(t) \right)^2 - \text{RSS}_{\text{clv},i} \right). \end{aligned} \tag{3.11}$$

Here, the residual sum of squares between $\hat{f}_i(t)$ and $f_i(t)$ are expressed relative to the RSS_i of the exponential fits $y_i(t)$ to the data. The subtraction of these constants does not affect training but helps to interpret the contributions to cost function. The targets are weighed such that for either the set of perfect targets (a singleton), single mismatches, or double mismatches, the weights w_i of all member targets i add to 1. Given the 90/10 train-test split of double-mismatched targets, the weights are

$$w_i = \begin{cases} 1 & \text{perfect target,} \\ 1/60 & \text{1 mismatch,} \\ 1/1539 & \text{2 mismatches (train).} \end{cases} \tag{3.12}$$

Optimization is performed with generalized simulated annealing [76], with initial temperature $T_0 = 300$, visiting parameter $q_V = 2.5$, and acceptance parameter $q_A = -5$. We optimize for 1000 iterations, in each of which the parameters are changed individually in 41 update steps. We carry out 200 independent training runs on the DelftBlue high-performance computer [77], and select the top-50% for further analysis. For optimization of the R-loop scaling factor α, we carry out 25 independent training runs at fixed values $\alpha = 0.5, 0.6, \dots, 1.0$ and obtain the expected optimum value by parabolic interpolation through the median fit scores.

3.4.6. Binding kinetics analysis

The Area Under the Curve (AUC) is calculated for binding fit curves $y_{\mathrm{bnd},i}(t)$ versus $\ln t$ in the experimental time frame $t \in [t_1, t_2]$,

$$\begin{aligned} \mathrm{AUC}_i &= \int_{\ln t_1}^{\ln t_2} y_{\mathrm{bnd},i}(t)\, d(\ln t) = \int_{t_1}^{t_2} \frac{y_{\mathrm{bnd},i}(t)}{t} dt \\ &= \phi_{\mathrm{bnd},i} \left(\ln\left(\frac{t_2}{t_1}\right) + E_1(k_{\mathrm{bnd},i}\, t_2) - E_1(k_{\mathrm{bnd},i}\, t_1) \right). \end{aligned} \tag{3.13}$$

The solution is expressed in terms of the exponential integral $E_1(x) = \int_x^\infty \exp(-x')\, dx'/x'$. The value of AUC_i is reported relative to its maximum value, $\ln(t_2/t_1)$.

3.4.7. Simulation of single-molecule FRET experiments

Aguirre Rivera et al. have provided us with the average FRET values for each of the 256 DNA library members in their assay [49]. The FRET histograms in their work follow a double Gaussian distribution, $p_L \mathcal{N}(\mu_L, \sigma_L) + p_H \mathcal{N}(\mu_H, \sigma_H)$, where the low and high FRET peaks are located at $\mu_L = 0.386$ and $\mu_H = 0.690$, and where p_L and p_H are the weights of the low and high FRET populations, defined such that $p_L + p_H = 1$. Accordingly, given the average FRET signal x_i on a DNA target i , we estimate the probability to attain the low FRET state by interpolation between μ_L and μ_H,

$$\hat{p}_{L,i} = \min\left(\max\left(\frac{\mu_H - x_i}{\mu_H - \mu_L}, 0 \right), 1 \right), \tag{3.14}$$

where the $\min$ and $\max$ operations guarantee probabilities in the range $\hat{p}_{L,i} \in [0, 1]$.

We select the best-scoring parameter sets from our library-1 trainings of CRISPRzip and the reference model. Correcting for the two point mutations in the gRNA sequence (U instead of C, at $b = 16$ and $b = 18$), free energy landscapes are generated for each of the DNA sequences. By assuming equilibrium over all states, we obtain the state occupancies p_b from Boltzmann statistics,

$$p_b = \frac{\exp(-U_b/k_B T)}{\sum_{m=0}^{20} \exp(-U_m/k_B T)}. \tag{3.15}$$

The low FRET state is chosen to correspond with R-loops of size 16, $P(b \geq 16) = \sum_{b=16}^{20} p_b$, because of the position of the Cy5 label at that position. Also, the requirement that $b \geq 16$ for low FRET signal results in a maximal correlation coefficient r for both the reference model and CRISPRzip predictions compared to the experimental data.

3.4.8. PAM-binding experiments

Perfect target 1 (**Table 3.S4**, Pr2.21, Azenta Life Sciences) and empty plasmid vector (**Table 3.S4**, Pl0.63) were PCR amplified (Phusion Plus polymerase, Thermo Scientific;

Table 3.S4, primers Pr6.29/Pr6.30 and Pr6.31/Pr6.32, Azenta Life Sciences) and then column purified (GeneJET PCR purification kit, Thermo Scientific). Perfect target 1 was inserted into empty plasmid vector via HiFi assembly (NEB) and electroporated (ECM630, BTX) into homemade *E. coli* cells (TOP10 competent cells, Invitrogen). Assembled "Perfect target 1" plasmid (**Table 3.S4**, Pl1.70) was purified from successful transformants (GeneJET plasmid miniprep kit, Thermo Scientific) and confirmed via sequencing (SeqVision UAB).

Two additional plasmids were made by site directed mutagenesis: one has a T instead of a C in position 3 of the target sequence (C3T), and the other has an A instead of a G in position 5 in Cas9 target sequence (G5A) (**Table 3.S4**). Primers with the desired substitutions (**Table 3.S4**, Pr7.77-Pr7.80) are used to PCR amplify DNA off the template plasmid (Phusion Plus Polymerase; **Table 3.S4**, Pl1.70). Each reaction was exposed to DpnI to remove template (Thermo Scientific) and then electroporated into *E. coli* cells (TOP10, homemade). Several randomly picked colonies were purified (GeneJET plasmid miniprep kit, Thermo Scientific) and their sequence was confirmed (SeqVision UAB). DNA targets – perfect target, C3T, G5A – were amplified from their plasmids via PCR (Phusion Plus Polymerase; **Table 3.S4**, Pr6.29, Pr6.30) and purified (GeneJET column purification kit, Thermo Scientific). Amplicon size and purity is confirmed with agarose gel electrophoresis in 1X TBE buffer (90 mM Tris-base, 90 mM boric acid, 2 mM EDTA, pH8).

The kinetics of PAM- binding and -unbinding were determined over a range of RNP concentrations (128 pM, 640 pM, 3.2 nM, 16 nM, 80 nM, 400 nM, 2 μM) This study used gRNA (**Table 3.S3**) made by GenScript, diluted in 1 mM Tris, 0.1 mM EDTA, pH 8.0 upon arrival. Efficient RNP assembly was ensured by mixing Cas9 with excess sgRNA 1 (2 μM, 2.4 μM, respectively; **Table 3.S3**) in r3.1 buffer, and incubating at 22 °C for 15 minutes. This RNP was also diluted to concentrations of 400 nM, 80 nM, and 16 nM. For the other concentrations, Cas9 was mixed with excess sgRNA (3.2 nM, 12.8 nM, respectively) in r3.1 buffer, and also diluted to RNP concentrations of 640 pM and 128 pM. Each RNP solution was combined with DNA at a final reaction concentration of 30 pM. The reactions were sampled and quenched periodically (0, 0.16, 0.5, 1, 3, 10, 30, 100 min) in stop buffer for 30 minutes at 37 °C. DNA cleavage was visualized on 6% Native PAGE and band intensity was quantified (GelAnalyzer 19.1, excluding background by rolling ball method, peak width tolerance = 20% of lane profile length).

3.4.9. DNA superhelicity effects on Cas9 dynamics

We estimate the work done by Cas9 to unwind the DNA helix on the basis of a single molecule data from Ivanov et al. (2020) [47]. In the study, magnetic tweezers twist a linear DNA substrate containing a target that perfectly or partially matches gRNA 1. A rotor bead reveals the transition rates k_{ij} between the R-loop states $i \in \{\mathrm{P}, \mathrm{I}, \mathrm{O}\}$, with "P" the PAM state, "I" the intermediate state, and "O" the open R-loop state. The torque τ follows from the total DNA twist ψ imposed by the magnetic tweezers, and the angle ω by which Cas9 locally unwinds DNA for R-loop formation, leading to overwinding of the tether. It is calculated as

$$\tau = \kappa \cdot (\psi + \omega), \tag{3.16}$$

with the parameters defined such that positive and negative ψ and τ respectively reflect over- and underwound DNA, and $\omega \geq 0$ by definition. The observed torsional stiffness of $\kappa = 1.48$ pN nm rot^{-1} (**Figure 3.S7**) is expected for a 5 kb tether with a twist persistence length $C = 100$ nm for linear DNA under a tension of 5 pN [78, 79]. R-loop completion corresponds to full B-DNA unwinding, $\omega = 681° \approx 20 \times 34°$ (**Figure 3.S7**).

In CRISPRzip, we assume constant DNA helix unwinding with angular distances φ to the transition states and θ between the R-loop states (**Figure 3.7A**). For DNA under a torque $\tau_0 = \kappa \cdot \psi$ in its unopened state, the free-energy landscape $U_b(\tau_0)$ can be obtained from the landscape $U_b(0)$ on unconstrained DNA by including the angular work to extend the R-loop,

$$\begin{aligned} U_b(\tau_0) &= U_b(0) + \int_0^{b\theta} \tau d\omega \\ &= U_b(0) + \tau_0 \cdot b\theta + \frac{1}{2}\kappa\,(b\theta)^2. \end{aligned} \tag{3.17}$$

Similarly, the R-loop extension rates $k_{b\to b+1}$ change according to the work to reach the next transition state,

$$\begin{aligned} k_{b\to b+1}(\tau_0) &= k_\mathsf{f} \cdot \exp\left(-\frac{1}{k_B T}\int_{b\theta}^{b\theta+\varphi} \tau d\omega\right) \\ &= k_\mathsf{f} \cdot \exp\left(-\frac{\tau_0 \cdot \varphi + \kappa\, b\theta\, \varphi + \frac{1}{2}\kappa\varphi^2}{k_B T}\right). \end{aligned} \tag{3.18}$$

If the DNA substrate has significant torsional stiffness κ, the R-loop extension rates are no longer constant across all states b.

We fix the value $\theta = 34°$ on the basis of the observed full B-DNA unwinding. The transition angle φ is estimated by fitting transition rates $\hat{k}_{ij}$ on perfect target 1 to experimental data (measured with twist cycling of 3 rpm and 15 rpm). For the PAM-intermediate transitions (k_{PI} and k_{IP}), kinetic data is added from the off-target with mismatches from position $b = 13$ on. To calculate $\hat{k}_{ij}$ in CRISPRzip, we 1) assume a position b of state "I"; 2) calculate the arrival time distributions between states "P", "I" and "O"; and 3) fit exponential functions to these distributions with fit rates $\hat{k}^b_{\mathsf{PI}}$, $\hat{k}^b_{\mathsf{IP}}$, $\hat{k}^b_{\mathsf{IO}}$ and $\hat{k}^b_{\mathsf{OI}}$. This process is carried out for assumed intermediate states $b \in \{9, \ldots, 13\}$, after which we determine the final effective rates according to Boltzmann statistics of the intermediate states,

$$\hat{k}_{ij} = \frac{\sum_{b=9}^{13} \exp(-U_b/k_B T) \cdot \hat{k}^b_{ij}}{\sum_{b=9}^{13} \exp(-U_b/k_B T)}. \tag{3.19}$$

We perform a log-least-squares fit of the transition angle φ to the experimental rates k_{PI}, k_{IP}, k_{IO} and k_{OI} as well as their ratios $k_{\mathsf{PI}}/k_{\mathsf{IP}}$ and $k_{\mathsf{IO}}/k_{\mathsf{OI}}$, the latter of which are added to assign extra weight to the relative state stabilities in the optimization.

In a genome, the torsion on DNA follows from the superhelical density σ, which quantifies how much the helix is overwound ($\sigma > 0$) or underwound ($\sigma < 0$) compared to its relaxed state. For an *E. coli* genome with $\sigma = -0.05$, the DNA helix makes 5%

fewer turns than it would at its natural helix repeat of $h_0 = 10.5\,\text{bp} = 3.6\,\text{nm}$. The torque on a genome is given by $\tau = C_{\text{eff}}\, 2\pi\, \sigma\, k_B T / h_0$, where $C_{\text{eff}} = 20\,\text{nm}$ is the effective twist persistence length of plectonemic, tensionless DNA [78, 79]. These values predict a torque of $\tau = -7.2\,\text{pN nm}$ on supercoiled bacterial genomes at $\sigma = -0.05$. Due to the large sizes of genomes (5 Mb for *E. coli*), their torsional stiffness is negligibly small ($\kappa \approx 10^{-4}\, k_B T \text{rot}^{-1}$) and a constant torque $\tau = \tau_0$ can be assumed.

3.5. Author contributions

H.S.O. and M.D. conceived and designed the study. H.S.O. developed the methodology, implemented the software, trained the model, validated the model, performed the analysis, and generated the visualizations. I.J. and S.K.J. developed the experimental protocols. I.J. performed the experiments. H.S.O. and S.K.J. curated the data. H.S.O. and I.J. wrote the original draft manuscript. H.S.O., I.J., S.K.J., and M.D. reviewed and edited the manuscript. M.D. supervised the project.

3.6. Acknowledgements

We thank Sebastian Deindl, Anton Sabantsev and Klaus Brackmann for their involvement in the project, for their experiments on salt conditions, and for supplying their MUSCLE data [49] and discussing how to model these. We thank Ralf Seidel, Julene Madariaga-Marcos and Fabian Welzel for our many valuable discussions about target interrogation dynamics and the implementation of nearest-neighbor models. We thank Ard Louis and Eryk Ratajczyk for our discussion on R-loop energetics and the influence of Cas9 on R-loop stability. We thank the EMBL sequencing facility (GeneCore, Heidelberg, DE) for sequencing assistance. Finally, we thank Raúl Ortiz-Merino, Elviss Dvinskis, and the TU Delft Digital Competence Centre for providing assistance in the design and development of the CRISPRzip Python package and the CRISPRzip GUI. H.S.O. acknowledges financial support from the Bionanoscience Department at TU Delft. This work was partially funded by the European Research Council (Project 101078247—PROTEGE; S.K.J.), the Ministry of Education, Science and Sports of the Republic of Lithuania (Program "University Excellence Initiatives" 12-001-01-01-01, project No. S-A-UEI-23-10; S.K.J.), and a European Molecular Biology Organization Installation Grant (Project IG 5728-2024; S.K.J.).

3.7. Data and software availability

The CRISPRzip source code (v1.2.2) [70] is available on GitHub (https://github.com/hiddeoff/crisprzip) and on PyPi (https://pypi.org/project/crisprzip/). The CRISPRzip GUI (v1.0.0) [69] is available on GitHub (https://github.com/hiddeoff/crisprzip-tool). The NucleaSeq pipeline for processing cleavage assay read counts [63] is available on Github (https://github.com/JonesLabEU/nucleaseq). The target sequences, experimental cleavage and binding kinetics, and CRISPRzip-simulated cleavage and binding kinetics for libraries 1 and 2 are provided as supplementary data to the preprint on bioRxiv (https://doi.org/10.64898/2025.12.12.691775). Raw sequencing data will be made available through the European Nucleotide Archive database.

3.8. Supplementary Figures

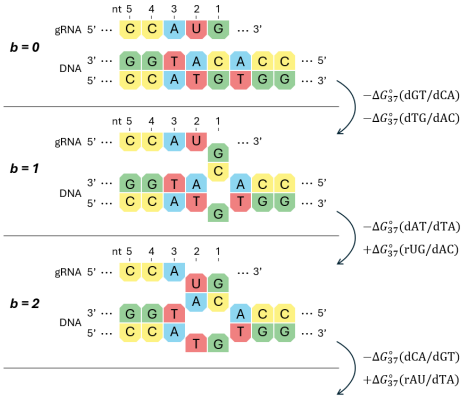


Figure 3.S1: Calculation of the R-loop cost. At each R-loop extension step, a DNA:DNA base pair is broken up and an RNA:DNA base pair is established. The nearest neighbor parameters describe the stability of a stack of two neighboring base pairs, e.g. the base stack 5'-GT-3' / 3'-CA-5' (or equivalently, 5'-AC-3' / 3'-TG-5') at position $b = 1$ has a free energy ΔG°_{37}(dGT/dCA). Alkan et al. (2018) [23] prescribes a collection of nearest-neighbor parameters and a methodology to add them, accounting for the presence of mismatches between the DNA and gRNA sequence. The total R-loop cost under standard conditions is calculated as a progressive sum over all nearest-neighbor parameters.

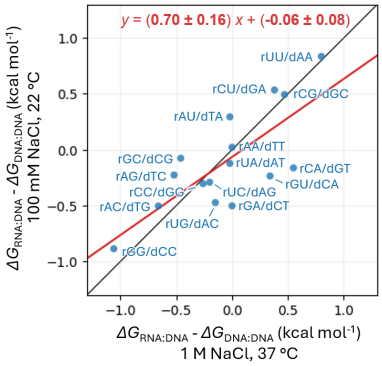


Figure 3.S2: Influence of temperature and salt on the base stack exchange cost. The base exchange cost $\Delta G_{\text{RNA:DNA}} - \Delta G_{\text{DNA:DNA}}$ represents the work required to replace a dsDNA base stack by a hybrid RNA:DNA base stack. No mismatching base pairs are included. The notation rGU/dCA indicates replacing 5'-GT-'3 / 3'-CA-5' dsDNA base pairs with 5'-GU-3' / 3'-CA-5' hybrid base pairs. Parameters at standard conditions (1 MNaCl, 37 °C) are from SantaLucia & Hicks (2004) [39] (dsDNA) and Sugimoto et al. (1995) [40] (RNA:DNA). Parameters at physiological salt (100 mM NaCl) and room temperature (22 °C) are from SantaLucia & Hicks (2004) [39] and Banerjee et al. (2020) [45].

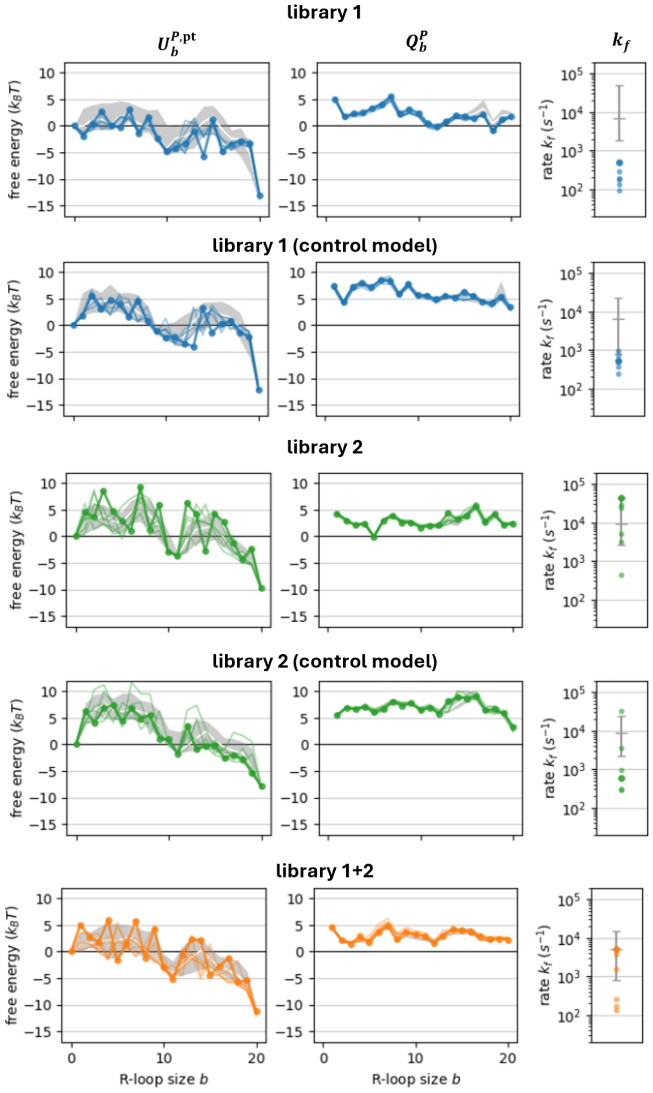


Figure 3.S3: Obtained parameter sets. Example parameter sets as obtained from different training settings. For each, 200 optimization runs are performed at fixed $\alpha = 0.7$. The parameter sets consist of 20 perfect-target protein contributions $U_b^{\mathrm{P,pt}}$, 20 protein mismatch penalties Q_b^{P} and the R-loop rate k_{f}. Solid, dotted line shows best-scoring parameter set, light lines show the 5 next-best scoring parameter sets (shown as large and small dots in the right panels). White line indicates median of the top 50% solutions, light gray region shows their 50% confidence intervals (shown as error bars in the right panels).

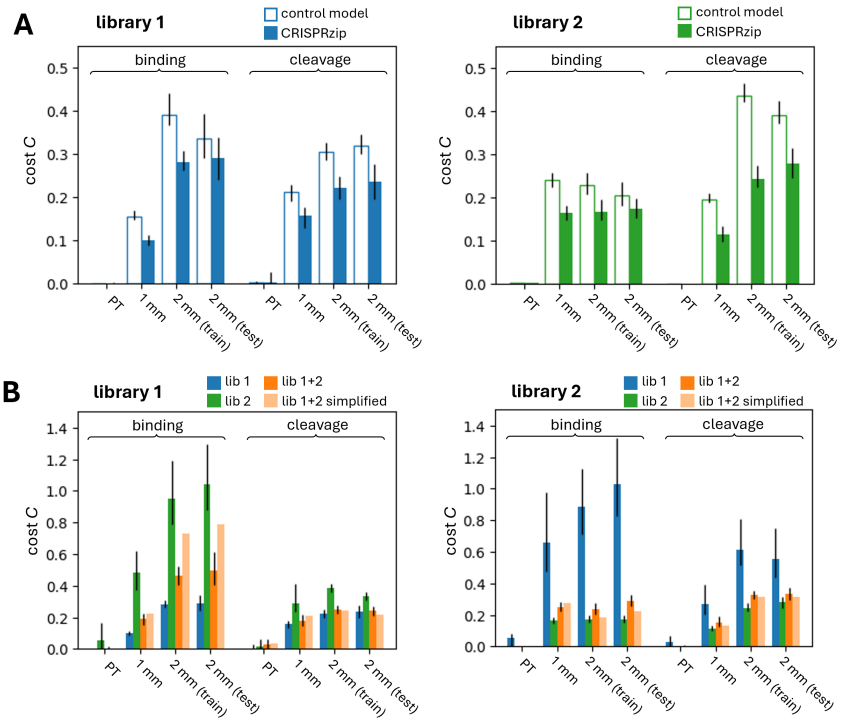


Figure 3.S4: Fit scores on data subsets. A Partial fit costs C on the perfect target (PT), single mismatch off-targets (1 mm) and double mismatch off-targets (2 mm), split according to the train and test group (90%/10% data split). Bars show average fit cost of the top 50% results from 200 optimization runs, error bars show 95% confidence interval. A sum of all partial fit costs C weighted according to their group size (**Equation 3.12**) results in the total fit score (**Figure 3.2B**). **B** Total and partial fit costs of the top-50% parameter sets from training on library 1 and 2 in isolation (blue, green) and simultaneously (orange), together with the performance of the simplified parameter set (**Figure 3.3B**).

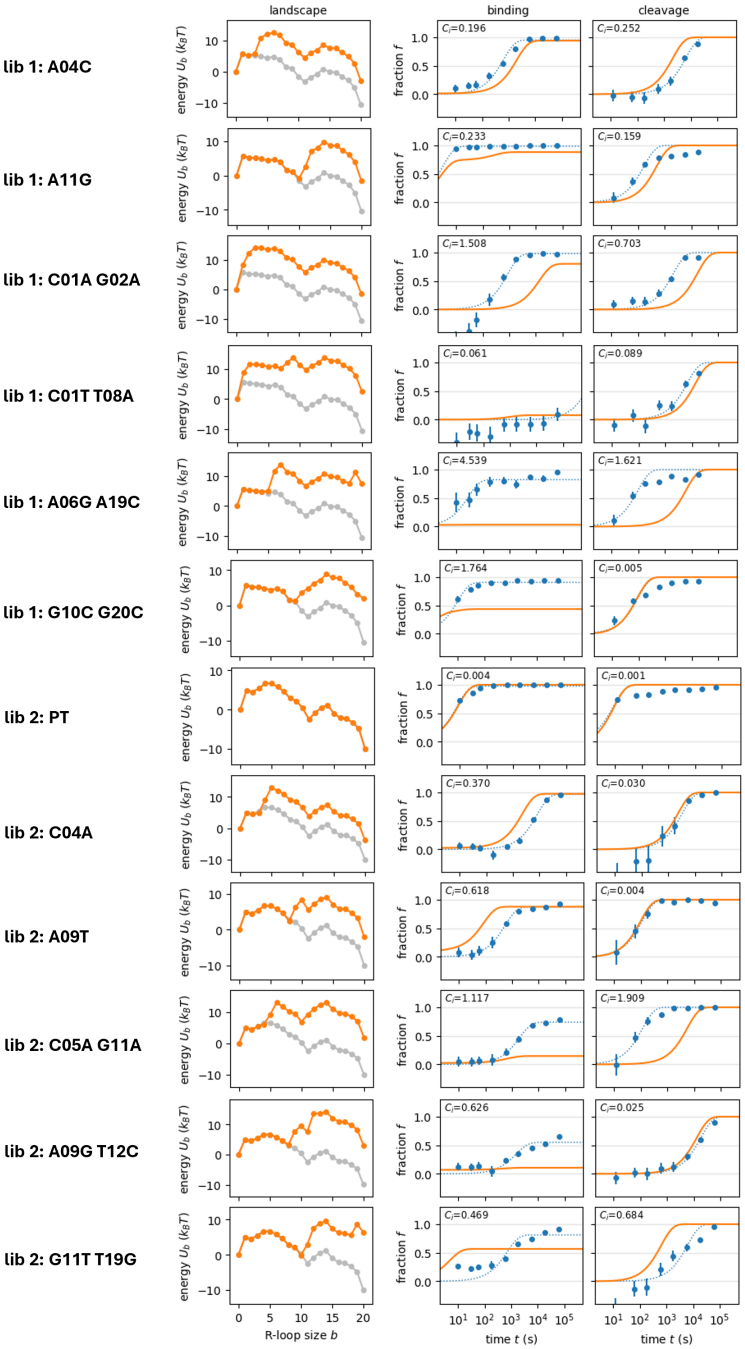


Figure 3.S5: CRISPRzip predictions Predicted landscapes and kinetics (orange) for several targets in libraries 1 and 2, made with the simplified parameter set. Perfect target landscapes (light gray) are shown for reference. Experimental data shown (blue) with error bars indicating standard deviation and with exponential fit curve (dotted line). Curve fit scores C_i reported next to binding and cleavage predictions.

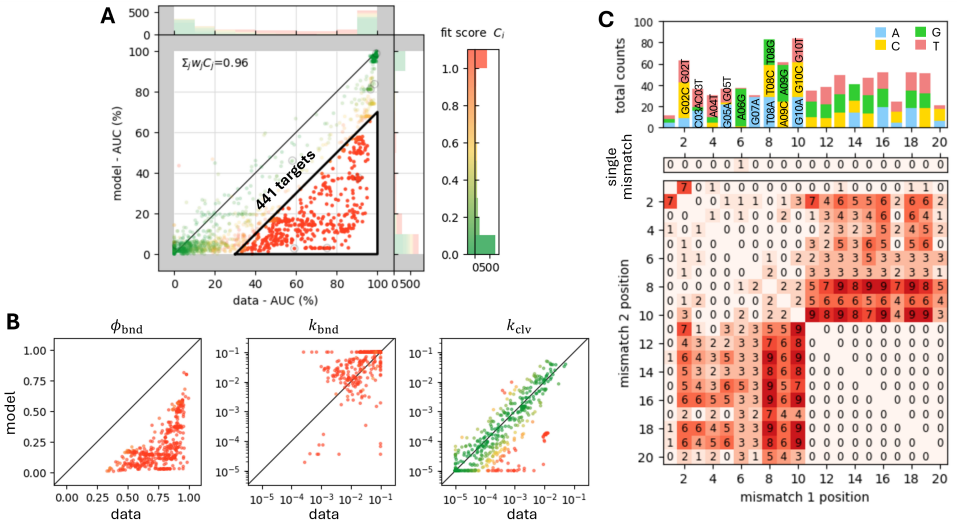


Figure 3.S6: Anomalously binding library 1-targets. A Selection of 441 targets in library 1 that display anomalously stable binding. **B** Correlating the parameters of exponential fit curves to the experimental and simulated data: the maximum bound fractions ϕ_{bnd}, binding rates k_{bnd} and cleavage rates k_{clv}. **C** Inspection of the sequences of the 441 targets. Top panel: total occurrence of specific mismatches, by base type. Center and bottom panel: occurrence of mismatch positions, in isolation (center) or in pairs (bottom). The selection includes at most 3 targets with a single mismatch at a particular position (center) and at most 9 targets with a mismatch pair at a particular position (bottom). See Figure 3.S5 for targets A06G A19C and G10C G20C.

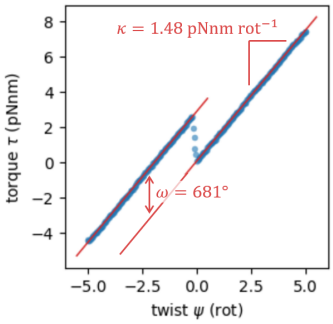


Figure 3.S7: Torsional stiffness in single-molecule twisting experiments. The torsional stiffness $\kappa = \tau/\psi$ of the full DNA tether is determined with a linear fit to the torque-twist curve.

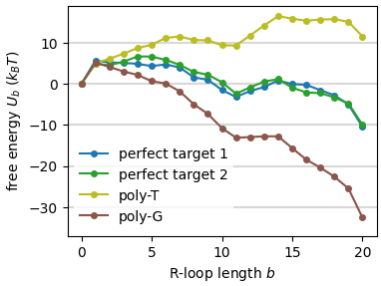


Figure 3.S8: Lower and upper bounds to R-loop stability. Free energy landscapes on the most stable (poly-G target) and least stable (poly-T target / poly-U gRNA) R-loops according to nearest-neighbor parameters. Target-1 and -2 landscapes are shown for reference.

3.9. Supplementary Data

parameter(s)	name	value	source
$U_1^{P,\text{pt}}, \dots, U_{20}^{P,\text{pt}}$	perfect-target protein contribution	free	
$Q_1^P, \dots, Q_{20}^P$	protein mismatch penalties	free	
$U_1^R, \dots, U_{20}^R$	R-loop cost	sequence-dependent	Alkan et al. (2018) [23]
α	R-loop scaling coefficient	0.7	This work
k_f	R-loop extension rate	free	
k_cat	catalysis rate	$4.0\,\text{s}^{-1}$	Gong et al. (2018) [37]
k_on^0	PAM-binding rate	$3.4 \cdot 10^{-2}\,\text{nM}^{-1}\,\text{s}^{-1}$	This work
k_off	PAM-unbinding rate	$1.0\,\text{s}^{-1}$	This work
θ	R-loop step angle	34°	Ivanov et al. (2020) [47]
φ	transition angle	50°	This work

Table 3.S1: Free and fixed parameters in CRISPRzip.

parameter	value (k_BT)	parameter	value (k_BT)	parameter	value (s^{-1})
$U_1^{P,pt}$	1.344	Q_1^P	4.445	k_f	292.6
$U_2^{P,pt}$	1.546	Q_2^P	2.097		
$U_3^{P,pt}$	1.766	Q_3^P	1.548		
$U_4^{P,pt}$	2.194	Q_4^P	2.048		
$U_5^{P,pt}$	1.990	Q_5^P	1.750		
$U_6^{P,pt}$	2.697	Q_6^P	3.657		
$U_7^{P,pt}$	2.065	Q_7^P	4.839		
$U_8^{P,pt}$	0.225	Q_8^P	2.566		
$U_9^{P,pt}$	-0.818	Q_9^P	3.237		
$U_{10}^{P,pt}$	-2.991	Q_{10}^P	2.816		
$U_{11}^{P,pt}$	-4.022	Q_{11}^P	2.703		
$U_{12}^{P,pt}$	-2.582	Q_{12}^P	1.851		
$U_{13}^{P,pt}$	-1.101	Q_{13}^P	3.164		
$U_{14}^{P,pt}$	0.193	Q_{14}^P	3.691		
$U_{15}^{P,pt}$	-1.401	Q_{15}^P	3.625		
$U_{16}^{P,pt}$	-2.889	Q_{16}^P	3.910		
$U_{17}^{P,pt}$	-3.552	Q_{17}^P	2.382		
$U_{18}^{P,pt}$	-4.464	Q_{18}^P	2.282		
$U_{19}^{P,pt}$	-6.045	Q_{19}^P	2.444		
$U_{20}^{P,pt}$	-11.154	Q_{20}^P	2.233		

Table 3.S2: Simplified parameter set.

Name	RNA sequence (5' to 3')
sgRNA 1	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUC CGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU GACGCAUAAAGAUGAGACGC
sgRNA 2	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUC CGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU GUGAUAAGUGGAAUGCCAUG

Table 3.S3: gRNA used in the study. Target-matching sequences are highlighted in red.

Name	DNA sequence (5' to 3')
Pr2.21 Perfect target 1	ATAACTAATTGAGCTGAACGCACCAACGTGGTACTCTGTGCA GATTTTAGACGCATAAAGATGAGACGCTGGTGATCGTCTTAG AGACAAGCACAGTCTCTCCACGAAGCATTACTCTAGCACAAC TACACGACTATTCAG
Pl0.63 Empty plasmid vector	ACGCGTCGCGAGGCCATATGGGTTAACCCATGGCCAAGCTT GCATGCCTGCAGGTCGACGGGCCCGGGAATGATACGGCGA CCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCTAACCGCCGAATAACAGAGTGTACAGTGTGCGAGGCGTT CTTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTTGTT CTTTTGCACTACCGTCAGGTAATCTCGTATGCCGTCTTCTGC TTGCCCGGGTTTAAACACATGTGTTTTTCCATAGGCTCCGCC CCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGG TGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCC CCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGC CGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGC GTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCG GTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACC CCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATC GTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTG GCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGT AGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACG GCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGA AGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCG GCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCA AGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGAT CCTTTGATCTTTTCTACGTTACGCCCCGCCCTGCCACTCATC GCAATACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAA GCCGTCACAGACGGCATGATGAACCTGAATCGCCAGCGGCA TCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGA AAACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAA TCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAA AAACATATTCTCAATAAACCCTTTAGGGAAATAGGCCAGGTT TTCACCGTAACACGCCACATCTTGCGAATATATGTGTAGAAA CTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATG...

Name	DNA sequence (5' to 3')
PI0.63 Empty plasmid vector (continued)	...AAAACGTTTCAGTTTGCTCATGGAAAACGGTGTAACAAGG GTGAACACTATCCCATATCACCAGCTCACCGTCTTTCATTGC CATACGGAATTCCGGATGAGCATTCATCAGGCGGGCAAGAA TGTGAATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTAC GGTCTTTAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTT ATAGGTACATTGAGCAACTGACTGAAATGCCTCAAAATGTTC TTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGT GATTTTTTTCTCCATACTCTTCCTTTTTCAATATTATTGAAGCA TTTATCAGGGTTATTGTCTCATGAGCGGATACATATGGGTT
PI1.70 Perfect target 1 plasmid	ACGCGTCGCGAGGCCATATGGGTTAACCCATGGCCAAGCTT GCATGCCTGCAGGTCGACGGGCCCGGGAATGATACGGCGA CCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCTATAACTAATTGAGCTGAACGCACCAACGTGGTACTCTGT GCAGATTTTAGACGCATAAAGATGAGACGCTGGTGATCGTCT TAGAGACAAGCACAGTCTCTCCACGAAGCATTACTCTAGCAC AACTACACGACTATTCAGAGATCGGAAGAGCACACGTCTGA ACTCCAGTCACTTGTTCTTTTGCACTACCGTCAGGTAATCTC GTATGCCGTCTTCTGCTTGCCCGGGTTTAAACACATGTGTTT TTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCG ACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAA GATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCT CCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTT TCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCT GTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTG GGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCG CCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGAC ACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATT AGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAA GTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTG GTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGA GTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAG CGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAA AAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGTTACG CCCCGCCCTGCCACTCATCGCAATACTGTTGTAATTCATTAA GCATTCTGCCGACATGGAAGCCGTCACAGACGGCATGATGA ACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCGT ATAATATTTGCCCATGGTGAAAACGGGGGCGAAGAAGTTGT CCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCC AGGGATTGGCTGAGACGAAAAACATATTCTCAATAAACCCTT TAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTT GCGAATATATGTGTAGAAACTGCCGGAAATCGTCGTGGTATT CACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAAA CGGTGTAACAAGGGTGAACACTATCCCATATCACCAGCTCA CCGTCTTTCATTGCCATACGGAATTCCGGATGAGCATTCATC AGGCGGGCAAGAATGTGAATAAAGGCCGGATAAAACTTGTG CTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATATCCAG...

Name	DNA sequence (5' to 3')
Pl1.70 Perfect target 1 plasmid (continued)	...CTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGA AATGCCTCAAAATGTTCTTTACGATGCCATTGGGATATATCAA CGGTGGTATATCCAGTGATTTTTTTCTCCATACTCTTCCTTTT TCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGC GGATACATATGGGTT
Pr6.32 HiFi forward (vector)	AGATCGGAAGAGCACACGTC
Pr6.31 HiFi reverse (vector)	AGATCGGAAGAGCGTCGTGT
Pr6.29 HiFi forward (insert)	ACACGACGCTCTTCCGATCTATAACTAATTGAGCTGAACGCA C
Pr6.30 HiFi reverse (insert)	GACGTGTGCTCTTCCGATCTCTGAATAGTCGTGTAGTTGTGC T
Pr7.77 SDM C3T forward	GACGCATAAAGATGAGATGCTGGTGATCGTC
Pr7.78 SDM C3T reverse	GACGATCACCAGCATCTCATCTTTATGCGTC
Pr7.79 SDM G5A forward	GACGCATAAAGATGAAACGCTGGTGATCGTC
Pr7.80 SDM G5A reverse	GACGATCACCAGCGTTTCATCTTTATGCGTC

Table 3.S4: Plasmids, DNA targets and their primers used in the study.

References

1. Wang, J. Y. & Doudna, J. A. CRISPR technology: A decade of genome editing is only the beginning. *Science* **379,** eadd8643. doi:10.1126/science.add8643 (2023).

2. Zhu, H., Li, C. & Gao, C. Applications of CRISPR–Cas in agriculture and plant biotechnology. *Nature Reviews Molecular Cell Biology* **21,** 661–677. doi:10.1038/s41580-020-00288-9 (2020).

3. Morshedzadeh, F. et al. An Update on the Application of CRISPR Technology in Clinical Practice. *Molecular Biotechnology* **66,** 179–197. doi:10.1007/s12033-023-00724-z (2024).

4. Chehelgerdi, M. et al. Comprehensive review of CRISPR-based gene editing: mechanisms, challenges, and applications in cancer therapy. *Molecular Cancer* **23,** 9. doi:10.1186/s12943-023-01925-5 (2024).

5. Zahedipour, F., Zahedipour, F., Zamani, P., Jaafari, M. R. & Sahebkar, A. Harnessing CRISPR technology for viral therapeutics and vaccines: from preclinical studies to clinical applications. *Virus Research* **341,** 199314. doi:10.1016/j.virusres.2024.199314 (2024).

6. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821. doi:10.1126/science.1225829 (2012).
7. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology* **20,** 698–714. doi:10.1038/s41580-019-0152-0 (2019).
8. Xue, C. & Greene, E. C. DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing. *Trends in Genetics* **37,** 639–656. doi:10.1016/j.tig.2021.02.008 (2021).
9. Zou, R. S. et al. Massively parallel genomic perturbations with multi-target CRISPR interrogates Cas9 activity and DNA repair at endogenous sites. *Nature Cell Biology.* doi:10.1038/s41556-022-00975-z (2022).
10. Moreb, E. A. & Lynch, M. D. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nature Communications* **12.** Number: 1, 5034. doi:10.1038/s41467-021-25339-3 (2021).
11. Zhang, H. et al. Deep sampling of gRNA in the human genome and deep-learning-informed prediction of gRNA activities. *Cell Discovery* **9,** 48. doi:10.1038/s41421-023-00549-9 (2023).
12. Cameron, P. et al. Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nature Methods* **14,** 600–606. doi:10.1038/nmeth.4284 (2017).
13. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nature Methods* **14,** 607–614. doi:10.1038/nmeth.4278 (2017).
14. Fu, R. et al. Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity. *Nature Communications* **13.** doi:10.1038/s41467-022-28028-x (2022).
15. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Research* **50,** 3616–3637. doi:10.1093/nar/gkac192 (2022).
16. Sharma, S. et al. A review on bioinformatics advances in CRISPR-Cas technology. *Journal of Plant Biochemistry and Biotechnology.* doi:10.1007/s13562-022-00811-3 (2022).
17. Chen, Y. & Wang, X. Evaluation of efficiency prediction algorithms and development of ensemble model for CRISPR/Cas9 gRNA selection. *Bioinformatics* **38,** 5175–5181. doi:10.1093/bioinformatics/btac681 (2022).
18. Zhang, G., Luo, Y., Dai, X. & Dai, Z. Benchmarking deep learning methods for predicting CRISPR/Cas9 sgRNA on- and off-target activities. *Briefings in Bioinformatics* **24,** bbad333. doi:10.1093/bib/bbad333 (2023).
19. Yuan, H. et al. An Overview and Comparative Analysis of CRISPR-SpCas9 gRNA Activity Prediction Tools. *The CRISPR Journal* **8,** 89–104. doi:10.1089/crispr.2024.0058 (2025).

20. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology* **17,** 148. doi:10.1186/s13059-016-1012-2 (2016).

21. Eslami-Mossallam, B. et al. A kinetic model predicts SpCas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity. *Nature Communications* **13.** doi:10.1038/s41467-022-28994-2 (2022).

22. Farasat, I. & Salis, H. M. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLOS Computational Biology* **12,** e1004724. doi:10.1371/journal.pcbi.1004724 (2016).

23. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biology* **19,** 1–13. doi:10.1186/s13059-018-1534-x (2018).

24. Zhang, D., Hurst, T., Duan, D. & Chen, S. J. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 8693–8698. doi:10.1073/pnas.1820523116 (2019).

25. Corsi, G. I. et al. CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context. *Nature Communications* **13.** Number: 1, 3006. doi:10.1038/s41467-022-30515-0 (2022).

26. Bisaria, N., Jarmoskaite, I. & Herschlag, D. Lessons from Enzyme Kinetics Reveal Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-Based Genome Editing. *Cell Systems* **4,** 21–29. doi:10.1016/j.cels.2016.12.010 (2017).

27. Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depken, M. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Reports* **22,** 1413–1423. doi:10.1016/j.celrep.2018.01.045 (2018).

28. Ratajczyk, E. J. et al. Controlling DNA–RNA strand displacement kinetics with base distribution. *Proceedings of the National Academy of Sciences* **122,** e2416988122. doi:10.1073/pnas.2416988122 (2025).

29. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573. doi:10.1038/nature13579 (2014).

30. Nishimasu, H. et al. Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* **156,** 935–949. doi:10.1016/j.cell.2014.02.001 (2014).

31. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348,** 1477–1481. doi:10.1126/science.aab1452 (2015).

32. Pacesa, M. et al. R-loop formation and conformational activation mechanisms of Cas9. *Nature* **609.** Number: 7925, 191–196. doi:10.1038/s41586-022-05114-0 (2022).

33. Rutkauskas, M. et al. Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Reports* **10,** 1534–1543. doi:10.1016/j.celrep.2015.01.067 (2015).

34. Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nature Communications* **7,** 1–8. doi:10.1038/ncomms12778 (2016).

35. Jiang, F. et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351,** 867–871. doi:10.1126/science.aad8282 (2016).

36. Zhu, X. et al. Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nature Structural & Molecular Biology* **26,** 679–685. doi:10.1038/s41594-019-0258-2 (2019).

37. Gong, S., Yu, H. H., Johnson, K. A. & Taylor, D. W. DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell Reports* **22,** 359–371. doi:10.1016/J.CELREP.2017.12.041 (2018).

38. Raper, A. T., Stephenson, A. A. & Suo, Z. Functional Insights Revealed by the Kinetic Mechanism of CRISPR/Cas9. *Journal of the American Chemical Society* **140,** 2971–2984. doi:10.1021/jacs.7b13047 (2018).

39. SantaLucia, J. & Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics* **33,** 415–440. doi:10.1146/annurev.biophys.32.110601.141800 (2004).

40. Sugimoto, N. et al. Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. *Biochemistry* **34,** 11211–11216. doi:10.1021/bi00035a029 (1995).

41. Sugimoto, N., Nakano, M. & Nakano, S.-i. Thermodynamics−Structure Relationship of Single Mismatches in RNA/DNA Duplexes. *Biochemistry* **39,** 11270–11281. doi:10.1021/bi000819p (2000).

42. Watkins Jr, N. E. et al. Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes. *Nucleic Acids Research* **39,** 1894–1902. doi:10.1093/nar/gkq905 (2011).

43. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research* **38,** D280–D282. doi:10.1093/nar/gkp892 (2010).

44. Nakano, S.-i., Fujimoto, M., Hara, H. & Sugimoto, N. Nucleic acid duplex stability: influence of base composition on cation effects. *Nucleic Acids Research* **27,** 2957–2965. doi:10.1093/nar/27.14.2957 (1999).

45. Banerjee, D. et al. Improved nearest-neighbor parameters for the stability of RNA/DNA hybrids under a physiological condition. *Nucleic Acids Research* **48,** 12042–12054. doi:10.1093/nar/gkaa572 (2020).

46. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67. doi:10.1038/nature13011 (2014).

47. Ivanov, I. E. et al. Cas9 interrogates DNA in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 5853–5860. doi:10.1073/pnas.1913445117 (2020).

48. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences* **114,** 5461–5466. doi:10.1073/pnas.1700557114 (2017).

49. Aguirre Rivera, J. et al. Massively parallel analysis of single-molecule dynamics on next-generation sequencing chips. *Science* **385,** 892–898. doi:10.1126/science.adn5371 (2024).

50. Newton, M. D. et al. Negative DNA supercoiling induces genome-wide Cas9 off-target activity. *Molecular Cell* **83,** 3533–3545.e5. doi:10.1016/j.molcel.2023.09.008 (2023).

51. Banerjee, T., Takahashi, H., Rendra, D., Subekti, G. & Kamagata, K. Engineering of the genome editing protein Cas9 to slide along DNA. *Scientific Reports* **11.** ISBN: 0123456789, 14165. doi:10.1038/s41598-021-93685-9 (2021).

52. Ratajczyk, E. J., Šulc, P., Turberfield, A. J., Doye, J. P. K. & Louis, A. A. Coarse-grained modeling of DNA–RNA hybrids. *The Journal of Chemical Physics* **160,** 115101. doi:10.1063/5.0199558 (2024).

53. Chen, Q. et al. Genome-wide CRISPR off-target prediction and optimization using RNA-DNA interaction fingerprints. *Nature Communications* **14,** 7521. doi:10.1038/s41467-023-42695-4 (2023).

54. Cofsky, J. C., Soczek, K. M., Knott, G. J., Nogales, E. & Doudna, J. A. CRISPR–Cas9 bends and twists DNA to read its sequence. *Nature Structural & Molecular Biology* **29,** 395–402. doi:10.1038/s41594-022-00756-0 (2022).

55. Bravo, J. P. K. et al. Structural basis for mismatch surveillance by CRISPR–Cas9. *Nature*. doi:10.1038/S41586-022-04470-1 (2022).

56. Boyle, E. A. et al. Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Science Advances* **7.** doi:10.1126/sciadv.abe5496 (2021).

57. Rutkauskas, M. et al. A quantitative model for the dynamics of target recognition and off-target rejection by the CRISPR-Cas Cascade complex. *Nature Communications* **13.** Number: 1, 7460. doi:10.1038/s41467-022-35116-5 (2022).

58. Aris, K. D. P. et al. Dynamic basis of supercoiling-dependent DNA interrogation by Cas12a via R-loop intermediates. *Nature Communications* **16,** 2939. doi:10.1038/s41467-025-57703-y (2025).

59. Shi, H. et al. Rapid two-step target capture ensures efficient CRISPR-Cas9-guided genome editing. *Molecular Cell* **85,** 1730–1742.e9. doi:10.1016/j.molcel.2025.03.024 (2025).

60. Kauert, D. J. et al. The energy landscape for R-loop formation by the CRISPR–Cas Cascade complex. *Nature Structural & Molecular Biology* **30.** Number: 7, 1040–1047. doi:10.1038/s41594-023-01019-2 (2023).

61. Pacesa, M. et al. Structural basis for Cas9 off-target activity. *Cell* **185,** 4067–4081.e21. doi:10.1016/j.cell.2022.09.026 (2022).

62. Leenay, R. T. et al. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell* **62,** 137–147. doi:10.1016/j.molcel.2016.02.031 (2016).

63. Jones, S. K. et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nature Biotechnology* **39.** ISBN: 4158702006465, 84–93. doi:10.1038/s41587-020-0646-5 (2021).

64. Lu, Q., Bhat, D., Stepanenko, D. & Pigolotti, S. Search and Localization Dynamics of the CRISPR-Cas9 System. *Physical Review Letters* **127,** 208102. doi:10.1103/PhysRevLett.127.208102 (2021).

65. Achar, Y. J., Adhil, M., Choudhary, R., Gilbert, N. & Foiani, M. Negative supercoil at gene boundaries modulates gene topology. *Nature* **577,** 701–705. doi:10.1038/s41586-020-1934-4 (2020).

66. Raguram, A., Banskota, S. & Liu, D. R. Therapeutic *in vivo* delivery of gene editing agents. *Cell* **185,** 2806–2827. doi:10.1016/j.cell.2022.03.045 (2022).

67. Karp, H. et al. Packaged delivery of CRISPR–Cas9 ribonucleoproteins accelerates genome editing. *Nucleic Acids Research* **53,** gkaf105. doi:10.1093/nar/gkaf105 (2025).

68. Moreb, E. A., Hutmacher, M. & Lynch, M. D. CRISPR-Cas "Non-Target" Sites Inhibit On-Target Cutting Rates. *The CRISPR Journal* **3,** 550–561. doi:10.1089/crispr.2020.0065 (2020).

69. Offerhaus, H. *CRISPRzip-tool* version v1.0.0. Dec. 2025. doi:10.5281/zenodo.17818523.

70. Offerhaus, H. *CRISPRzip* version 1.2.2. Dec. 2025. doi:10.5281/zenodo.17818249.

71. Zhang, Z., Lamson, A. R., Shelley, M. & Troyanskaya, O. Interpretable neural architecture search and transfer learning for understanding CRISPR–Cas9 off-target enzymatic reactions. *Nature Computational Science* **3,** 1056–1066. doi:10.1038/s43588-023-00569-1 (2023).

72. Knott, G. J. & Doudna, J. A. CRISPR-Cas guides the future of genetic engineering. *Science* **361,** 866–869. doi:10.1126/science.aat5011 (2018).

73. Huang, X., Yang, D., Zhang, J., Xu, J. & Chen, Y. E. Recent Advances in Improving Gene-Editing Specificity through CRISPR–Cas9 Nuclease Engineering. *Cells* **11.** Number: 14, 2186. doi:10.3390/cells11142186 (2022).

74. Collias, D. & Beisel, C. L. CRISPR technologies and the search for the PAM-free nuclease. *Nature Communications* **12,** 555. doi:10.1038/s41467-020-20633-y (2021).

75. Hawkins, J. A., Jones, S. K., Finkelstein, I. J. & Press, W. H. Indel-correcting DNA barcodes for high-throughput sequencing. *Proceedings of the National Academy of Sciences* **115,** E6217–E6226. doi:10.1073/pnas.1802640115 (2018).

76. Tsallis, C. & Stariolo, D. A. Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications* **233,** 395–406. doi:10.1016/S0378-4371(96)00271-3 (1996).

77. Delft High Performance Computing Centre (DHPC). *DelftBlue Supercomputer (Phase 2)* https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2. 2024.

78. Marko, J. F. in *Mathematics of DNA Structure, Function and Interactions* 1st ed. *The IMA Volumes in Mathematics and its Applications* 150, 225–249 (Springer, New York, NY, 2020). ISBN: 978-1-4419-0669-4.

79. Gao, X., Hong, Y., Ye, F., Inman, J. T. & Wang, M. D. Torsional Stiffness of Extended and Plectonemic DNA. *Physical Review Letters* **127,** 028101. doi:10.1103/PhysRevLett.127.028101 (2021).

4

Supercoiling twists Cas9 off-target discrimination when binding, nicking and cleaving

Ieva Jaskovikaitė, Hidde S. Offerhaus, Marius Vinogradovas, Ursule Barkauskaitė, Martin Depken, Stephen K. Jones

Programmed with an RNA guide, Cas9 directs double-stranded DNA cleavage via its two nuclease domains. However, Cas9 sometimes falsely identifies DNA targets by binding and cleaving DNA that does not match its PAM and gRNA sequence requirements. Cas9's specificity is often affected by DNA topology, as DNA negative supercoiling can increase off-target activity while positive supercoiling can even prevent on-target activity. Such dramatic DNA topological changes routinely occur in cells as a result of transcription and replication, making Cas9's specificity a challenge for gene editing. To determine how Cas9 imparts its specificity across sequences and topologies, we directly mapped kinetic cleavage rates and sites for thousands of relaxed and negatively-supercoiled target sequences via NucleaSeq. We find that negative supercoiling accelerates off-target cleavage up to a thousand-fold, and can shift cleavage sites by two nucleotides. Guide-target mispairs differentially impacts RuvC and HNH domains, which can lead to topology-dependent nicking by Cas9. Finally, we predict these variations in Cas9 cleavage activity with a biophysical model that accounts for DNA topological state. These efforts expose Cas9's strand-specific off-target cleavage activity and will improve off-target identification for more predictable and safer gene editing.

H.S.O. contributed to this chapter through modeling and data analysis; all experimental work was carried out by other authors. See Author contributions (**Section 4.5**).

This chapter is under preparation for submission. A preprint is available on bioRxiv.

4.1. Introduction

The *Streptococcus pyogenes* Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)–associated 9 (Cas9) nuclease forms a ribonucleoprotein (RNP) complex with a guide RNA (gRNA; made of CRISPR-RNA and trans-activating RNA) [1], that is capable of recognizing target DNAs [2]. Upon locating an NGG protospacer adjacent motif (PAM), Cas9 forms an R-loop by melting the adjacent DNA and pairing it with its gRNA [2–4]. Given sufficient complementarity, the two nuclease domains, HNH and RuvC, cleave target and non-target DNA strands, respectively [5, 6]. Due to simple gRNA reprogramming of Cas9 to different targets, the nuclease is widely used in the lab and attractive for the clinic [7–9].

Cas9 occasionally operates on partially-matched targets, which can trigger edits at unintended locations and even chromosomal rearrangements [10, 11]. Such lack of specificity includes accepting non-canonical PAM sequences (i.e. NGA and NAG) [12, 13] and mispairs within the 20-nucleotide target [14–20]. Cas9 tolerates gRNA-relative mispairs *in vitro* and *in vivo* [14–20], with greater tolerance for PAM-distal mispairs than for PAM-proximal "seed" mispairs (positions ~5–10) [17]. Cas9's cumulative activity across off-targets can be as frequent as it is for on-targets [15, 16, 21]. For gRNAs tested on the human genome, single base mismatches are most prevalent within off-targets [21], even within off-targets containing other mispairs (gRNA-relative deletions or insertions) [15]. Concerns about off-targets have motivated investigations into off-target binding patterns [14, 22], cleavage patterns [14, 16], and their relation to gRNA sequence [23–25]. These studies, and the models built upon them [26–32], help make genome editing safer by pinpointing, predicting and avoiding Cas9's promiscuity. Unfortunately, current *in silico* prediction models still miss *bona fide* off-targets, and would benefit from including additional variables that impact Cas9 activity.

Recent studies show that DNA topology—specifically negative supercoiling (nSC)—can increase Cas9 off-target cleavage [16, 33, 34], and processes like transcription induce such topological changes [35]. Negative supercoiling induces transient DNA unwinding [36, 37] and increases Cas9's propensity for cleaving genomic off-targets [16]. Here, negative torque on the DNA stabilizes R-loop intermediates for genomic off-targets with PAM-distal mispairs, encouraging cleavage. However, PAM-proximal mispairs discourage R-loop intermediate formation, and negative torque provides insufficient assistance to overcome them [33]. Similarly, catalytically "dead" Cas9 (dCas9) can dwell longer on off-targets that are underwound [33, 34]. Yet, we still miss a clear understanding of the role that supercoiling plays across Cas9 binding and cleavage in the wider off-target space (in relation to mispair types, identities and positions), rather than mismatch mispairs alone [38]. We also understand that Cas9 cuts relaxed targets at different positions (and even trims ends) depending on target sequence [6, 14, 39], but similarly miss whether supercoiling changes cleavage positions within off-targets. This is especially critical, as changes in cleavage position alter editing outcomes with Cas9 [40].

In this high-throughput study, we investigate how supercoiling impacts where and how quickly Cas9 cleaves off-targets. Our results offer genome-agnostic, nucleotide-resolution kinetics and DNA end analysis for supercoiled libraries of systematically-mispaired off-targets. By expanding application of the NucleaSeq method

[14] to supercoiled targets, we directly relate our results to our previous study on relaxed DNA. We show that Cas9's ability to cleave targets depends on the mispairs present and the target's topology. Supercoiling increases Cas9's cleavage kinetics for off-targets with seed-region mispairs, but cleaves off-targets with PAM-distal mispairs similarly whether supercoiled or not. We observe that active Cas9 either nicks or repositions cleavage of some off-targets with PAM-proximal indel mispairs. Finally, we use our biophysical CRISPRzip model (**Chapter 3**) to rationalize the effects we observed in supercoiled target experiments. The model allows us to predict Cas9 cleavage kinetics across target sequences and supercoiling levels. Together, our data and model show the range of Cas9 activity users can expect across the DNA topologies found in our genome.

4.2. Results

4.2.1. NucleaSeq captures Cas9 cleavage kinetics on supercoiled DNA

The NucleaSeq method measures nuclease cleavage kinetics and cleavage position for relaxed, linear DNAs [14]. However, negative supercoiling (nSC) plays an important role in Cas9 targeting specificity [16, 33, 34], so we adapted NucleaSeq to capture its effect on nuclease specificity during cleavage (**Figure 4.1A**). Our libraries [14, 41] included targets that perfectly matched Cas9's PAM and gRNA sequences gRNA1 and gRNA2 (as used in [14, 22, 33, 42–45]) those with alternative PAMs, non-target control sequences, and thousands of off-target sequences with gRNA-relative single or double mismatches, insertions and deletions (**Figure 4.1B**). Each target sequence was flanked by unique barcodes for post-cleavage identification.

We inserted each of our libraries' members into a minimal 1.6kb plasmid vector and transformed them into *E. coli*. Purified plasmid libraries were negatively supercoiled, with a superhelical density of $\sigma = -0.033 \pm 0.007$ (**Figure 4.S1**) [46]. For nSC cleavage kinetics via NucleaSeq, we exposed each plasmid library to the Cas9 ribonucleoprotein complex (RNP) at 22 °C, periodically quenching samples from the reaction (**Methods 4.4**). After removing the plasmid vector (**Figure 4.1D**), we appended time-associated barcodes (e.g. sequencing indices) during preparation for next generation sequencing (NGS). Samples were pooled together and sequenced (**Figure 4.1C**). Sequencing read counts for each library member and timepoint were normalized (as described in Jones et al. (2021) [14]) and fit to exponential functions to determine effective cleavage rates (**Figure 4.1E,F**).

4.2.2. Mispairs affect Cas9 target cleavage rate and position

To investigate Cas9 specificity across targets when cleaving nSC DNA, we investigated off-targets with three mispair types – mismatches, insertions or deletions (when comparing target DNA to the gRNA used; **Figure 4.2**, **4.3**). Mismatches can involve non-canonical base pairing, while insertions and deletions can introduce nucleotide bulges and base skipping [47]. Our previous NucleaSeq results used the same libraries and analysis as present [14], enabling direct comparisons between relaxed and nSC DNA cleavage kinetics. Relaxed DNA off-targets experience very slow cleavage by

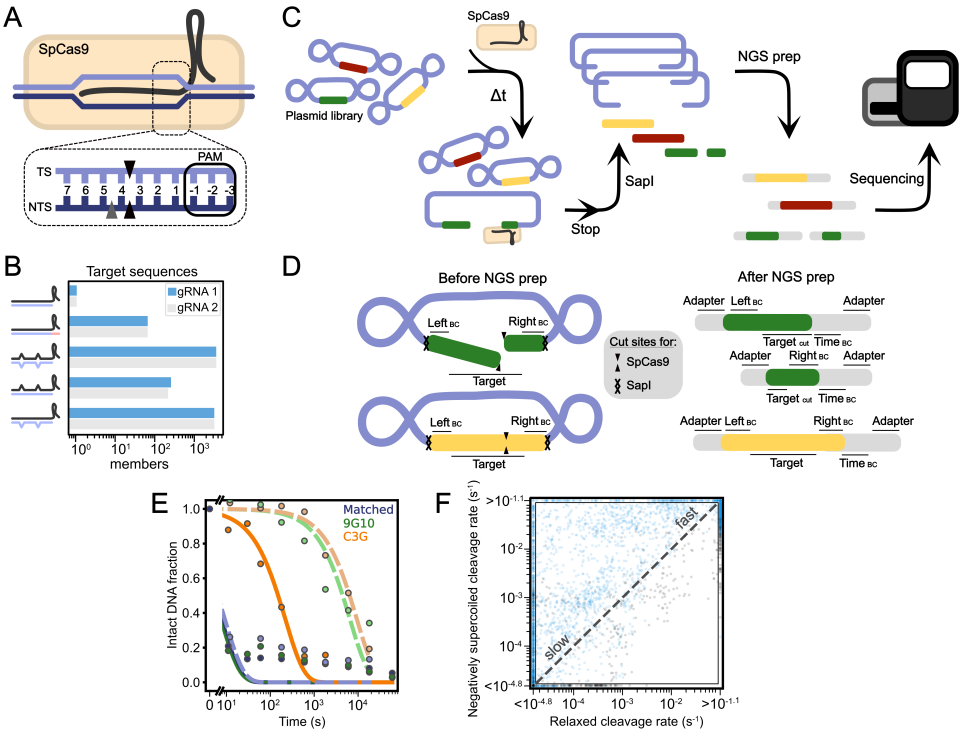


Figure 4.1: Adapting NucleaSeq for plasmid cleavage rate and site determinations. A Cas9 typically cleaves on-target DNA between the 3rd and 4th bases of the target (TS) and non-target (NTS) strands (black arrows). Some on-targets experience non-target strand (NTS) cleavage between the 4th and 5th bases (grey arrow). **B** Composition of targets in plasmid libraries for gRNA 1 and 2. Both libraries include fully-matched (gRNA and PAM) on-targets, off-targets with alternative PAMs, and off-targets with one or two mismatches, insertions and deletions relative to the gRNA. **C** The updated NucleaSeq workflow assays plasmid cleavage. Plasmid library is exposed to Cas9 for increasing times (Δt = 0, 0.16, 0.5, 1, 3, 10, 30, 100, 300, 1000 min) before quenching (Stop). The quenched reactions contain cleaved and intact targets. Then, samples are restricted with SapI and the plasmid vector is removed from library targets based on size. DNAs remaining in each sample receive a time barcode during NGS preparation, and are finally sequenced by short-read NGS. **D** A diagram shows individual cut (green) and intact (yellow) DNAs from a supercoiled target library before (left) and after (right) NGS prep. Initially, two unique barcodes (Left BC, Right BC) and SapI sites (hatch) flank each target. Cas9 cleavage (arrows indicate site) separate one side's unique barcode and SapI site from the other side's. Whether cut or intact, each piece of DNA receives a time barcode (Time BC) during the adapter ligation step of NGS preparation. **E** Fraction of DNA remaining intact upon Cas9 exposure over time for gRNA and PAM-matched on-target (Matched, purple) and two off-targets (9G10, green; C3G, orange). Read counts (circles; intense colors=nSC while matched pastel colors=Rel) are normalized at each time point to non-target controls, and then to the t=0 read counts. Exponential functions are fit to read normalized counts from supercoiled DNA (solid lines) or relaxed DNA (dashed lines). **F** Library-wide comparisons of cleavage rates for supercoiled versus relaxed target DNA. Blue: Library 1 for gRNA 1. Grey: Library 2 for gRNA 2. Dashed line shows x=y.

Cas9 when containing PAM-proximal “seed” mispairs as opposed to PAM-distal mispairs (**Figure 4.2**). Here, Cas9 cleaved these same off-targets much faster when negatively supercoiled, regardless of whether mispairs were proximal or distal, in agreement with a prior study [38], and across our target libraries (**Figure 4.2**). Cas9 cleaved nSC off-targets with single mismatches in the seed region up to 100-fold faster than relaxed counterparts, while rates for most PAM-distal nSC off-targets were so fast as to be indistinguishable from the on-target rate. Surprisingly, off-targets with a single deletion mispair in the first 4 PAM-proximal positions were cleaved slower when supercoiled than relaxed. Such unpaired bases may misalign the target DNA with Cas9’s cleavage domains, and supercoiling could exacerbate this misalignment [47]. Off-targets with insertions at positions 6–10 showed much faster cleavage on supercoiled DNA than relaxed DNA, but not so for off-targets with insertions at positions 11–17—a region where hybridization licenses Cas9 cleavage [48]. Among single mispair off-targets, the difference between nSC and relaxed DNA cleavage rates was greatest for insertion mispairs: some were cleaved >10k-fold faster when supercoiled. For some off-targets with multiple (2+) mispairs, Cas9 still cleaves them, in agreement with cell-based experiments [18]. These results highlight that Cas9 effectively “queries” DNA topology when discriminating off-targets.

Finally, we analyzed differences in Cas9 cleavage sites library-wide, with respect to supercoiling. Cas9 typically cleaves on-target DNAs between the third and fourth nucleotides from the PAM, or occasionally the fourth and fifth on the non-target strand (depending on the gRNA) [3, 14]. Similarly, mispairing and supercoiling might alter the orientation of each DNA strand to its nuclease domain, thereby shifting cleavage sites. We found that nSC off-targets with single mismatches were largely cleaved at the same positions as their relaxed counterparts, and similarly to the on-target (**Figure 4.3**) [14]. However, for off-targets with PAM-proximal deletions, Cas9 often cleaved one nucleotide closer (on the target strand; TS) or further (on the non-target strand; NTS) from the PAM, independent of supercoiling. Off-targets with single PAM-proximal insertions showed the greatest topology sensitivity: upon supercoiling, target strand cleavage shifted as much as two nucleotides further from the PAM. This may result from downstream R-loop hybridization more strongly impacting nucleotide positioning than PAM-PID interactions. We investigated whether this extended to off-targets with two PAM-proximal insertions, and found that indeed for some, TS cleavage sites were shifted an additional nucleotide further away. Thus, while Cas9 slowly cleaves off-targets with PAM-proximal mispairs, they nonetheless experience the greatest shift in cleavage sites relative to supercoiling.

4.2.3. Cleavage site-adjacent mispairs disrupt Cas9 nuclease activity

Mispairing between the gRNA and target strand DNA distorts the heteroduplex structure within Cas9 [49]. Supercoiled off-targets (with PAM-proximal insertion mispairs) can experience Cas9 cleavage further from the PAM (**Figure 4.3C**) indicating that supercoiling further distorts the heteroduplex. We hypothesized that this distortion impacts cleavage by Cas9’s RuvC and HNH nuclease domains differently, which cannot be detected via dsDNA cleavage alone. We selected three off-targets for deeper investigation: “C3G”, “Δ3”, and “3A4” (DNA with G-C replacing C-G at position 3, no

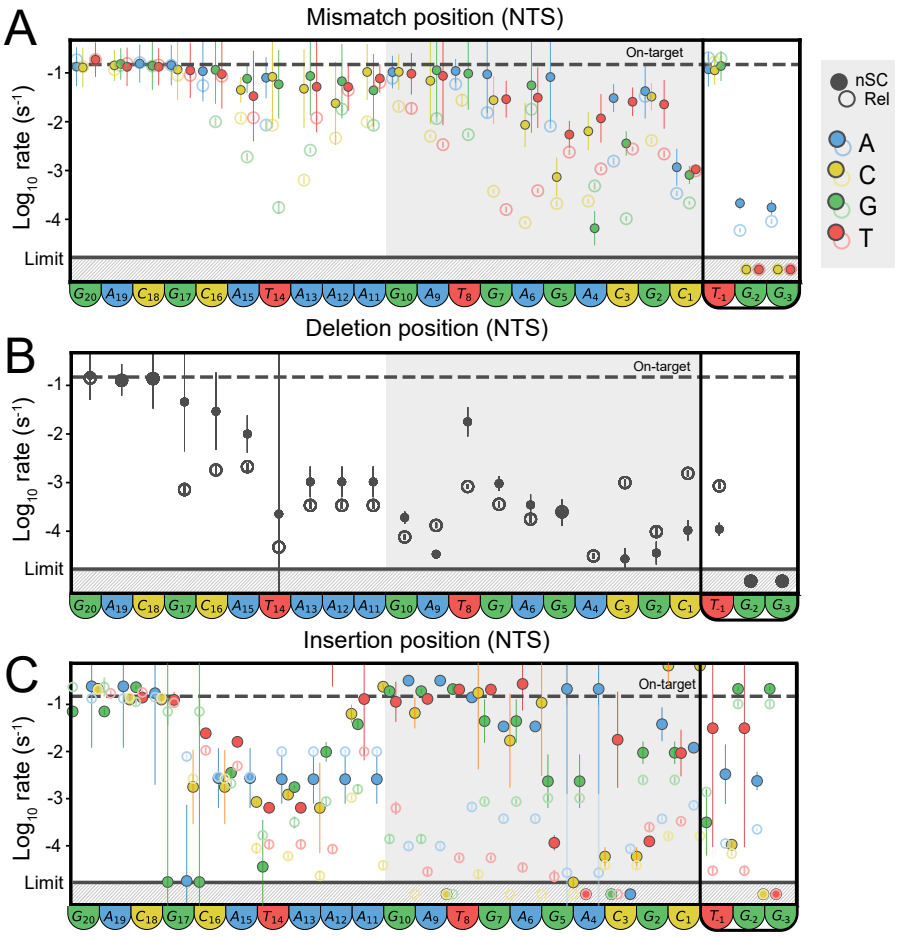


Figure 4.2: Supercoiling affects Cas9 cleavage rates in a mispair-dependent manner. Fitted Cas9 cleavage rates for targets with one PAM- or gRNA1-relative mismatch (**A**), deletion (**B**), or insertion (**C**). Hollow circles: relaxed target values. Filled circles: supercoiled target values. Targets with fewer than 30 reads at t=0 were omitted. Error bars: SD based on 50 rounds of bootstrapping with replacement. X axis labels: NTS nucleotides and positions in relation to the PAM and gRNA1. Grey box: approximate seed region. Dashed lines: on-target cleavage rate (nSC and Rel). Solid lines: detection limits. Hatched regions contain fit values exceeding the adjacent detection limit.

position 3 base pair, or an extra A-T between base pairs 3 and 4, with respect to the gRNA-matched DNA sequence) (**Figure 4.3**). For Δ3 and 3A4, Cas9 nicked—but did not cut—the DNA (**Figure 4.4A,B**). This suggests that at least one nuclease domain loses its cleavage capacity where “extra” DNA or gRNA bases distort local pi-stacking and structure, a phenomenon not observed with just a single mismatch [50] (**Figure 4.4A,B**).

To pinpoint which Cas9 cleavage domain becomes impaired—HNH or RuvC, we

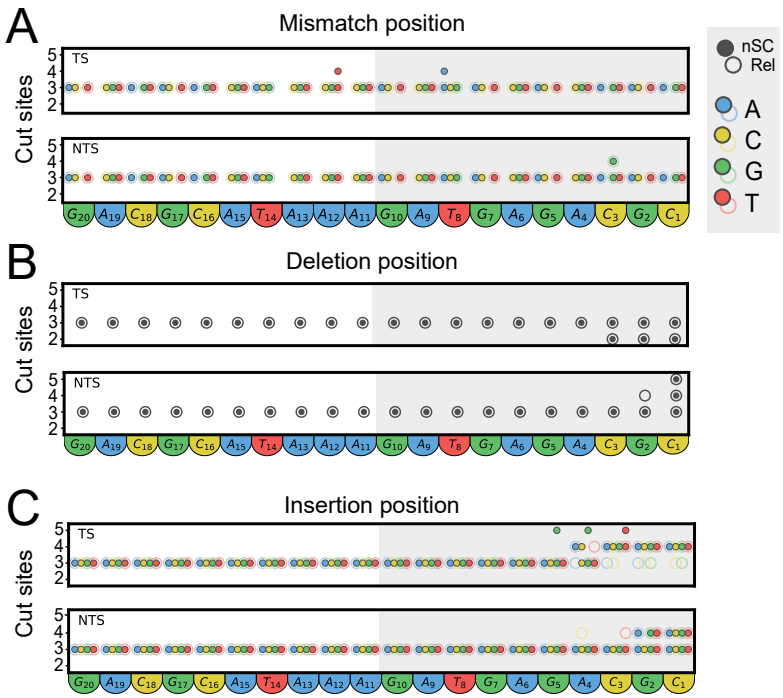


Figure 4.3: Supercoiling affects Cas9 cleavage positions in a mispair-dependent manner. Cas9 cleavage positions for targets with one PAM- or gRNA1-relative mismatch (**A**), deletion (**B**), or insertion (**C**). Supercoiled (filled circles) or relaxed (hollow circles) target cleavage positions are shown for the target strand (TS, top) and non-target strand (NTS, bottom). Indicated cleavage positions were found in ≥10% of cut reads for one or more timepoints. Cleavage positions are calculated away from the PAM. X axis labels: NTS nucleotides and positions in relation to the PAM and gRNA1.

quantified single and double nicking events by WT Cas9 and HNH- or RuvC-active only Cas9 nickases on supercoiled Δ3 and 3A4 off-targets (**Figure 4.4C–F**). Then, we measured HNH and RuvC behavior individually by performing these experiments with relaxed, fluorescently-labeled DNAs (**Figure 4.4C–F**).

Wild-type Cas9 (WT Cas9) only nicked nSC Δ3 (**Figure 4.4C**) off-target plasmid once, with no detectible second nick activity (**Figure 4.4D**). When investigating this nSC plasmid with nickase RNPs, only the RuvC-active Cas9 generated a nick. This suggested that WT Cas9's RuvC operates on this target, but its HNH does not. When investigating Cas9 activity with fluorescently labelled DNA (**Figure 4.4C**), both domains generated cleavage products, albeit only faintly for HNH (**Figure 4.4D**). This activity agrees with Cas9's known topology-independent nicking activity [38, 50, 51]. When investigating our off-target with an unpaired gRNA nucleotide at the typical cleavage site (target 3A4, **Figure 4.4E**), we observed topology-dependent nicking: HNH activity only occurred when the target was supercoiled (**Figure 4.4F**). Thus, a target's sequence and topological state together convert Cas9 from a nuclease to a nickase.

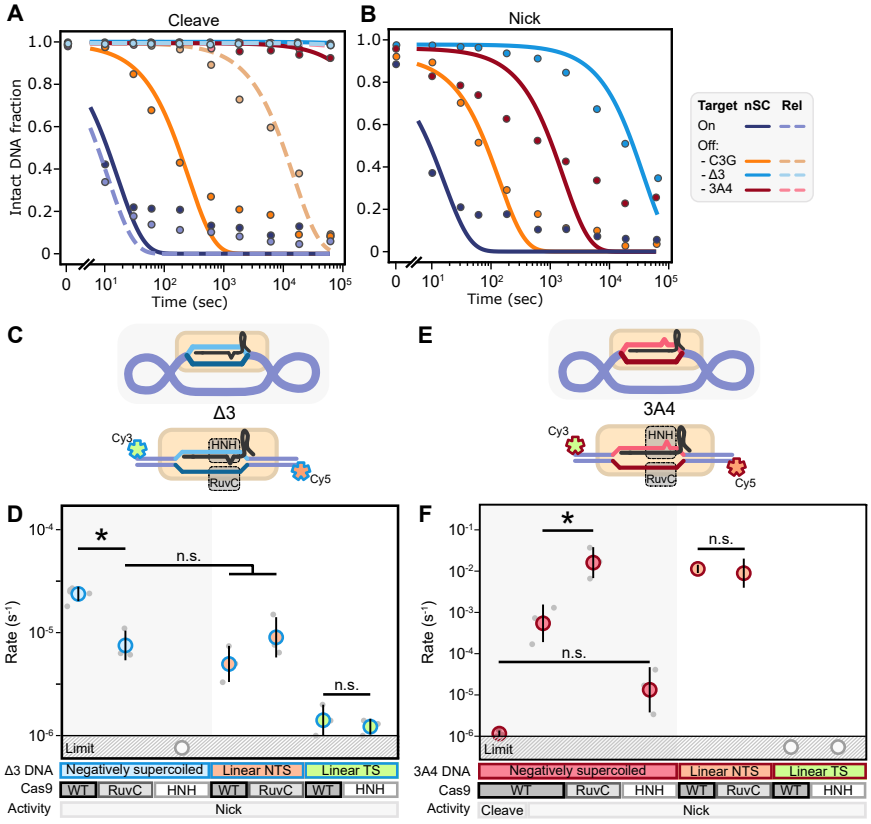


Figure 4.4: Cleavage site-adjacent mispairing differentially inhibits Cas9 nuclease domains. Cas9 cleavage (**A**) and nicking (**B**) of supercoiled or relaxed plasmids containing off-targets with mispairs at position 3. Single exponential functions were fit to lane-normalized agarose gel band intensities for supercoiled DNA (solid lines) or relaxed DNA (dashed lines). N=1. **C** Schematics for off-target Δ3 DNAs used in D. **D** Fit nicking rates for off-target Δ3 (gRNA1-relative deletion) when present in supercoiled plasmid DNA or relaxed linear DNA. WT Cas9, ΔH840A Cas9 nickase (active RuvC) or D10A Cas9 nickase (active HNH) were used as indicated. Single exponential functions were fit to lane-normalized agarose gel band intensities for all samples with detectable cut DNA in at least the last two timepoints (300 min and 1000 min). Relaxed linear DNAs were co-labeled with a Cy5 fluorophore on the non-target strand (NTS) and a Cy3 fluorophore on the target strand (TS). Solid line: detection limit. Mean±SD. N≥3. * p≤0.05 with Tukey's post-hoc test, after testing for normality. **E** Schematics for off-target 3A4 DNAs used in F. **F** Fit cleavage or nicking rates for off-target 3A4 (gRNA1-relative insertion) when present in supercoiled plasmid DNA or relaxed linear DNA. WT Cas9, ΔH840A Cas9 nickase (active RuvC indicated) and D10A Cas9 nickase (active HNH indicated) were used. Single exponential functions were fit to lane-normalized agarose gel band intensities for all samples with detectable cut DNA in at least the last two timepoints (300 min and 1000 min). Relaxed linear DNAs were individually labeled with a Cy5 fluorophore on the non-target strand (NTS) or a Cy3 fluorophore on the target strand (TS). Solid line: detection limit. Mean±SD. N≥3. * p≤0.05 with Tukey's post-hoc test, after testing for normality.

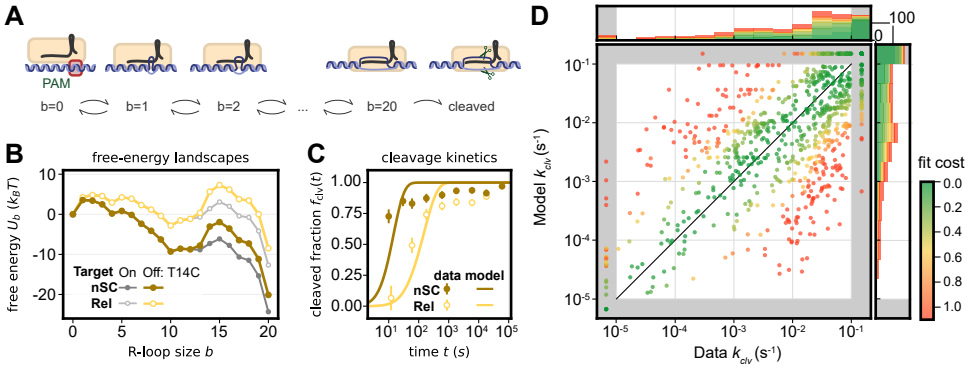


Figure 4.5: Modeling the mechanistic effects of supercoiling. **A** The CRISPRzip model simulates the kinetics of R-loop formation at base pair resolution. At state b, where b hybrid DNA:RNA base pairs are established, the free-energy difference with the neighboring states $b+1$ and $b-1$ respectively determine the rates of R-loop extension and recession. **B** Modeled free-energy landscapes of R-loop formation for off-target T14C DNA when relaxed (yellow) or supercoiled (brown). Free-energy landscapes for the on-target are shown for reference (grey). Negative supercoiling effectively tilts the landscapes down. **C** Cleavage of target T14C, as measured experimentally (circles) and as predicted with CRISPRzip (solid lines). The negative landscape tilt results in accelerated cleavage. Error bars: SD based on propagated uncertainty from Poisson-distributed initial read counts. **D** Measured vs CRISPRzip-predicated cleavage rates k_{clv}, after training CRISPRzip on the library-1 nSC cleavage data, which resulted in a fitted plasmid torque of $\tau_0 = -3.6 \pm 1.1$ pN nm. Circles: data points for targets from library 1. Colors indicate fit cost of that target.

4.2.4. A biophysical model captures the impact of supercoiling on Cas9 specificity

To obtain physical insight into our observed effects of supercoiling on Cas9 specificity (**Figure 4.1**, **4.2**), we applied our mechanistic "CRISPRzip" model (**Chapter 3**). CRISPRzip incorporates the effects of torque on Cas9 activity as previously demonstrated on single molecules [33]. This model leverages the fact that Cas9's R-loop forms progressively during target interrogation (**Figure 4.5A**). CRISPRzip simulates the kinetics of this process as stochastic progress through sequence-dependent free-energy landscapes (**Figure 4.5B,C**). We obtained the underlying landscape parameters by training on (d)Cas9 (binding) cleavage dynamics for relaxed targets with up to 2 point mutations (library-1, **Figure 4.S2**, **Methods 4.4**). We integrated the effect of torque [33] by letting DNA underwinding reduce the work needed to extend the R-loop, effectively tilting R-loop landscapes downward and accelerating cleavage (**Figure 4.5B,C**). DNA overwinding does the opposite.

To assess if our high-throughput kinetic data can be explained mechanistically, we trained the torque level in CRISPRzip to the nSC cleavage kinetics measured across library 1, as it contained the largest target set (**Figure 4.1B**). The obtained estimate for the plasmid torque ($\tau_0 = -3.6 \pm 1.1$ pN nm) corresponds to plasmids with superhelical density $\sigma = -0.03 \pm 0.01$ (**Chapter 3**), which agrees well with our supercoiling estimates from gel electrophoresis (**Figure 4.S1**). Visualizing the predicted

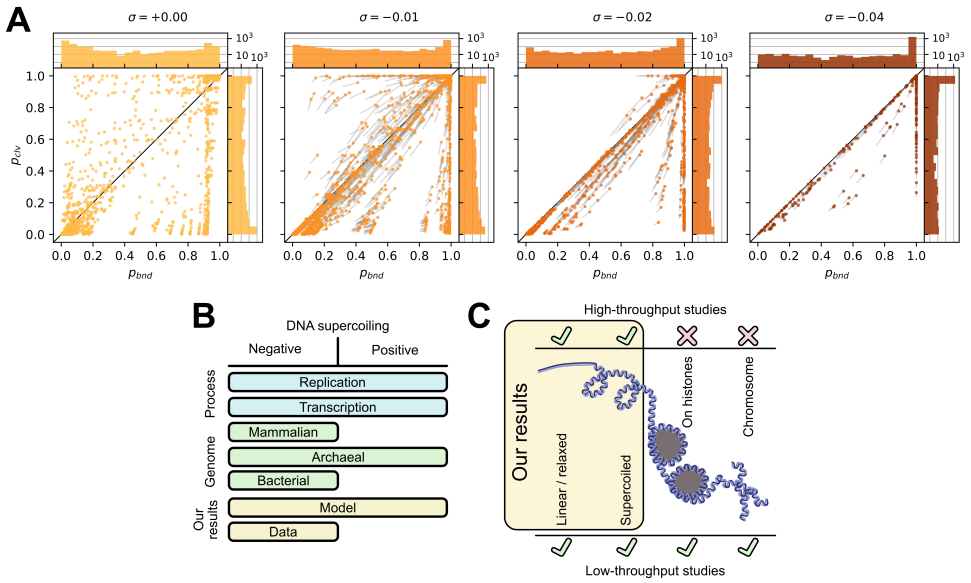


Figure 4.6: CRISPRzip predicts the mispair-dependent role of supercoiling in Cas9 binding and cleavage. A CRISPRzip predicts the probability that dCas9 binds (p_{bnd}) and Cas9 cleaves (p_{clv}) all targets with 0–2 mismatches after 15 minutes of exposure to 20 nM RNP. Tails (gray) show each target's sensitivity to supercoiling as changes in p_{bnd} and p_{clv} over $\sigma + 0.001$ to σ, where σ is the superhelical density. **B** CRISPRzip predicts Cas9 activity for negative and positive DNA supercoiling, as found across genomes and cellular processes. **C** High-throughput *in vitro* studies provide greater resolution and context to results from lower-throughput studies that describe how DNA organization impacts Cas9 specificity and utility for genome editing.

and measured cleavage rates shows that CRISPRzip reproduces the cleavage kinetics of most targets, spanning the full dynamic range (**Figure 4.5D**). The remaining variation may be due to sequence-specific responses to supercoiling that are outside the model scope.

We next leveraged CRISPRzip to understand the relationship between Cas9 off-target binding and cleavage upon increased negative supercoiling. While binding is necessary for cleavage to occur, even transient binding suffices. Weakly-bound off-targets may be cleaved, just as well-bound targets may not be (**Chapter 3**). CRISPRzip modeling indicates that "fast-cleaved but weakly-bound" targets become more stably bound with increased negative supercoiling (**Figure 4.6A**; targets in the top left clear sooner upon nSC). Further, CRISPRzip predicts the impact that positive supercoiling has on Cas9 activity (**Chapter 3**)—as seen during replication, transcription and in archaeal genomes (**Figure 4.6B**).

4.3. Discussion

Here, we investigated Cas9 off-target discrimination in a genome-agnostic manner, exposing its intrinsic binding and cleavage specificity, and its dependence on target supercoiling. We used the biophysical model CRISPRzip for a kinetic understanding of (d)Cas9's supercoiling sensitivity. Our work complements existing studies, which typically investigate Cas9 specificity in cellular environments or on extracted genomes (human, mouse, etc.) [52, 53]. Such results capture the impact of unique genome characteristics like chromatin accessibility and transient topologies on Cas9 specificity. Thus, off-target models built on these results exchange robust application across genomes for accuracy within the genome used for training [52, 54]. Indeed, differential gene expression across cell cycle phases and types dramatically impacts editing efficiency, editing outcomes and off-target activity [53, 55, 56]. Our use of synthesized libraries profiles Cas9 specificity with challenging yet comprehensive target sets that exceed any individual genome's contents.

Supercoiling represents a basic, nearly-universal aspect of topological DNA organization, yet is the highest topological level for which high-throughput biochemical Cas9 specificity studies are available (**Figure 4.6C**). The next step is compaction upon nucleosomes. Available *in vitro* studies have examined extremely tight nucleosome wrapping, which effectively prevents Cas9 access [57–59]. However, natural nucleosome breathing (and chromatin remodeling) reveals targets, where cleavage rates become equivalent to naked DNA targets [59, 60]. Further, Cas9 dwells on cellular targets for several minutes after cleavage [61] but experiences dislocation by other factors [62], thereby operating in a multiple-turnover regime. Thus, understanding how nucleosome positioning and movement impacts Cas9 activity and specificity remains a critical, but largely under-defined, factor that limits predictable action in the cell.

Target-gRNA mispair tolerance is also observed across other CRISPR-Cas enzymes: CRISPR-Cas12a cleaves off-targets with mispairs and non-canonical PAMs [14, 63], as does Cas12b, Cas12j, and Cas9 variants. For Cas12a, R-loop propagation and cleavage rates across off-targets strongly correlate, as its mechanism relies heavily on full R-loop propagation for successful dsDNA cleavage [64]. Even so, Cas12a shows promiscuous and accelerated cleavage for supercoiled targets [65–67], while Cascade shows very low binding and cleavage activity on relaxed targets [68–70]. Within the Cas9 family, for archaeal AtCas9 and AhCas9, supercoiling nearly abolishes PAM requirements for supercoiled targets, but not for relaxed ones [71]. Enzymes from other defense systems, like Gabija, also show preference for supercoiled DNA for its cleavage [72]. Thus, Cas9's mispair tolerance reflects a typical, rather than exceptional, feature found in defense systems.

Our results show two "proofs-of-concept" where Cas9 could be developed: as a supercoiling detector, or a mutation-free nickase. We reported that some off-targets experience a massive increase in Cas9 cleavage rate when negatively supercoiled (e.g. 9G10; **Figure 4.3**, **Figure 4.S3**). By programming Cas9 with a gRNA that recreates such a mispair, one would effectively convert DNA cleavage into a sensor for DNA torque. This could offer a simple and targeted alternative to intercalator assays [46]. We also reported that mispairing (3A4, Δ3; **Figure 4.4D,F**) can suppress HNH domain activity, converting WT Cas9 into a nickase. Base and prime editing are both enhanced

by nicking, where prime editing uses HNH-inactive Cas9 [73, 74]. Thus, mispairing gRNAs may offer reinforcement, or even replacement, of inactivating mutations in these applications.

4.4. Methods

4.4.1. sgRNAs

All sgRNAs used in this study were purchased from Synthego or GenScript and diluted in 1 mM Tris, 0.1 mM EDTA, pH 8.0 (**Table 4.S1**).

4.4.2. Plasmid library construction

This study used barcoded target libraries (GenScript) [14, 41] amplified by PCR (Phusion Plus polymerase, Thermo Scientific, **Table 4.S2**, primers Pr6.29/Pr6.30 for library 1; Pr7.26/Pr7.27 for library 2) with 25 cycles or less to maintain library diversity and then column purified (GeneJET PCR purification kit, Thermo Scientific). Plasmid backbone (**Table 4.S2**, Pl0.63) was similarly amplified with 20 cycles or less with Pr6.31/Pr6.32 for library 1 and Pr7.28/Pr7.29 for library 2 (**Table 4.S2**), then column purified. Both libraries were inserted into empty plasmid backbone via HiFi assembly (NEB), using a 5:1 ratio of library to backbone.

4.4.3. Plasmid library transformation and extraction

HiFi-assembled libraries were electroporated (ECM630, BTX; 1.8kV) into TOP10 *E. coli* cells (Invitrogen). To ensure efficient transformation and sufficient library representation, 104-fold dilutions were plated on Lb Agar + chloramphenicol (25µg/mL) and grown at 37 °C for 20h. Target libraries 1 and 2 had ≥120 and ≥90 colonies per library member collected from one or more transformations. Transformed cells were plated on 3+ 150mm Lb agar plates with chloramphenicol (25µg/mL). After growing the cells at 30 °C for 20-22h, the plates were scraped with 3mL LB media a total of three times, and then purified (Pure II Plasmid Midiprep kit, Zymo Research, D4201).

4.4.4. Superhelical density determination

To determine the superhelical density of purified plasmids, a previous protocol was modified [46] to accommodate our compact plasmid (1.7 kb). A subset of purified plasmids was relaxed with Topoisomerase I (NEB). Supercoiled and relaxed plasmids (0.5µg each) were ran on a 1% agarose gel (TBE, 2.4µg/mL of chloroquine diphosphate (Fisher Scientific) in the gel and the buffer) at 3V/cm for 4 hours, then post-stained with SYBR Safe (Invitrogen). We calculated band intensity profiles using GelAnalyzer 19.1. The change in linking number distribution (ΔLk) was calculated by counting bands from the middle of relaxed DNA to the middle of supercoiled DNA bands (**Figure 4.S1**), and then superhelical density was calculated as described in Martínez-García et al. (2018) [46].

4.4.5. NucleaSeq

The reaction components—sgRNA (**Table 4.S1**, Cas9 (NEB), plasmid library—were diluted in 100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl2, 100 µg mL−1 recombinant albumin, pH=7.9 (r3.1 buffer, NEB), to the final reaction concentrations of 187.5nM, 62.5nM, 6.25nM, respectively. First, diluted sgRNA and Cas9 were mixed to form an RNP for 15 minutes at 22°C. After the addition of the plasmid library, small aliquots of the reaction were removed at different timepoints—0, 0.2, 0.5, 1, 3, 10, 30, 100, 300 and 1000 min—and added to the STOP solution (60mM EDTA, Carl Roth; 1U Proteinase K, Thermo Scientific), then incubated at 37 °C for 30 minutes. Timepoint 0 was treated with STOP solution before the addition of DNA. Each timepoint sample was then ethanol precipitated.

The cleaned timepoint samples had their backbones cut out with an excess of SapI (NEB). The resulting reactions were measured with High Sensitivity Assay (Qubit) and half of each reaction was used for appending time indexes with NGS prep (NEBNext, NEB). Barcoded samples were size-selected with AMPure XP magnetic beads (Beckman Coulter) using 0.9X bead:DNA ratio, ran on capillary electrophoresis (BioAnalyzer 2100, Agilent) and then sequenced on a NextSeq2000 (P2 100bp PE or P3 100bp PE chemistry) at the EMBL GeneCore (Heidelberg, DE) with 10% PhiX spike to increase sequence diversity.

4.4.6. NGS data analysis: identification, normalization and rate fitting

Read counts for each library member were tabulated as previously described [14]. Briefly, paired NGS reads were merged, then filtered by quality and size. Acceptable reads for intact DNAs required correct matching to both barcodes and the target for a single library member, while cut DNAs required correct matching to a single barcode and the portion of the remaining target. After assigning reads to their matching library member, all read counts were scaled to the median non-target read count for each time point. For library members with 30 initial read counts ($t = 0$), read counts for each timepoint were normalized to the initial read counts. Standard deviations in read counts were obtained by assuming Poisson-distributed read counts and propagating errors. The experimental floor (defined as the median final timepoint value for all on-target controls) was first subtracted from all normalized read counts. Next, these adjusted normalized read counts were fit to a single exponential function to determine a cleavage rate for each library member i ($k_{\text{clv},i}$): $y_{\text{clv},i}(t) = \exp(-k_{\text{clv},i} \cdot t)$. Error in the fit was estimated by 50 rounds of bootstrapping with replacement.

4.4.7. Single targets construction

Target sequences (Azenta, **Table 4.S3** primers Pr2.21, Pr6.77, Pr6.79, Pr7.09, Pr7.32) were amplified with Phusion Plus polymerase (**Table 4.S2** primers Pr6.29 and Pr6.30). Vector (**Table 4.S2** Pl0.63) was the same for each target and was PCR amplified (Phusion Plus polymerase, **Table 4.S2** primers Pr6.31 and Pr6.32). All reactions were column-purified (GeneJet column purification kit). Targets Δ3 and 3A4 had Cy3 and Cy5 labels added via PCR (Phusion Plus polymerase, **Table 4.S3** primers Pr7.50 and

Pr8.27). Δ3 had both Cy3 and Cy5 simultaneously, while 3A4 had either one added at a time. Cy3 was always placed on a target strand and Cy5 on a non-target strand. All reactions were column-purified (GeneJet column purification kit).

Plasmids were assembled using HiFi assembly (NEB, 5:1 insert to vector ratio) and electroporated (TOP10 competent cells, homemade). The plasmids were extracted from several randomly picked colonies (GeneJet plasmid miniprep kit, Thermo Scientific; Monarch plasmid miniprep kit, NEB) and sent out for sequencing (SeqVision UAB). Colonies with the correct assemblies were amplified (ZymoPURE II plasmid midiprep kit, Zymo Research).

Target C3G was made from "perfect" on-target plasmid (**Table 4.S3**, PI1.70) by site-directed mutagenesis (Phusion Plus polymerase, 56 °C annealing, **Table 4.S3** primers Pr8.54 and Pr8.55). The resulting reaction was treated with DpnI (Thermo Scientific) and electroporated (TOP10 competent cells, homemade). The plasmids were extracted from several randomly picked colonies (GeneJet plasmid miniprep kit, Thermo Scientific; Monarch plasmid miniprep kit, NEB) and sent out for sequencing (SeqVision UAB). Colonies with the correct assemblies were amplified (ZymoPURE II plasmid midiprep kit, Zymo Research). Half of each plasmid sample was relaxed with Topoisomerase I (NEB) or Nb. BsrDI (NEB). The relaxation was confirmed on a 1% Agarose gel electrophoresis in TBE buffer and the plasmids were then ethanol purified.

4.4.8. Single target nicking and cleavage assays and analysis

The assays were performed on three types of DNA topologies for each of the selected single targets—supercoiled plasmid, relaxed plasmid, and linear PCR product. The assay was performed with Cas9 (NEB), gRNA (**Table 4.S1**, R11) and DNA diluted in 100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl2, 100 µg mL−1 recombinant albumin, pH=7.9 (r3.1 buffer, NEB). After the RNP was formed by incubating Cas9 with an excess of gRNA for 15 minutes at room temperature, DNA was exposed to it for different amounts of time (0, 0.2, 0.5, 1, 3, 10, 30, 100, 300, ≥1000 min) before the reaction was stopped (60mM EDTA, 1U Proteinase K), and incubated at 37 °C for 30 minutes. Some additional replicates omitted two final timepoints. The cleavage products were then visualized on 1% agarose (plasmid cleavage and nicking), 10% native PAGE (linear DNA cleavage) or 12% Urea PAGE in TBE buffer (linear DNA nicking).

To obtain cleavage rates from the gels, each gel was analyzed with GelAnalyzer 19.1 to quantify DNA band intensity. For fluorescently labelled targets, we quantified bands (~200–180 bp) and cut products (~90–68bp) after subtracting bands (~90–85bp) present in the control. The parameters used on the program included background removal by a rolling ball with a peak width tolerance in 20% of lane profile length. Then, the values were normalized per lane and timepoint, and scaled to $t = 0$. Finally, we obtained cleavage rate $k_{\mathrm{clv},i}$ by fitting a single exponential function to the diminishing population of intact or nicked DNA: $y_{\mathrm{clv},i}(t) = \exp(-k_{\mathrm{clv},i} \cdot t)$. We similarly obtained nicking rate $k_{\mathrm{nck},i}$ by fitting a function to the diminishing population of intact DNA only: $y_{\mathrm{nck},i}(t) = \exp(-k_{\mathrm{nck},i} \cdot t)$.

4.4.9. Statistics

After obtaining cleavage rates, each set of single target assay replicates were tested for normal distribution using Shapiro-Wilk test, and the significance between tested conditions were analyzed using One-Way ANOVA and Tukey's multiple comparison post-hoc test with Past v4.03 [75]. If no cleaved products appeared in the last two timepoints, no cleavage rate was assigned to that target.

4.4.10. Model methods

We fitted the mechanistic model CRISPRzip to the cleavage kinetics in library 1, as follows: first, we established a parameter set for this library on linear DNA, following the procedure described in **Chapter 3**. The landscape parameters ($U_b^{\mathrm{P,pt}}$, Q_b^{P} and k_{f}) were trained to the relaxed-DNA cleavage and binding kinetics for all 1770 targets with up to 2 mismatches in 200 independent training runs. From the 50% best-scoring results, we calculated the median protein contributions ($U_b^{\mathrm{P,pt}}$ and Q_b^{P}), and determined the corresponding R-loop extension rate (k_{f}) by optimization (**Figure 4.S2**). Next, we estimated the plasmid torque level τ_0 on the basis of the CRISPRzip torque parameters $\theta = 34°$ and $\phi = 50°$ estimated from torque spectroscopy [33] (**Chapter 3**). This torque-aware CRISPRzip version was trained to all supercoiled library-1 targets with ≥50 initial counts (1/1 perfect target, 60/60 single-mismatch targets, 851/1709 double-mismatch targets). Our modeling explicitly considered the coupling between the total R-loop opening angle ω and the plasmid torque τ,

$$\tau = \tau_0 + \kappa \cdot \omega \tag{4.1}$$

where τ_0 is the initial plasmid torque ($\tau_0 < 0$ for negative supercoils), and κ is the plasmid's torsional stiffness. For our 1.7-kb plasmids (length $L = 578\,\mathrm{nm}$) with effective torsional persistence length $C_{\mathrm{eff}} = 20\,\mathrm{nm}$ [76] we used $\kappa = C_{\mathrm{eff}}/L\,k_BT = 0.035\,k_BT\,\mathrm{rot}^{-1} = 0.9\,\mathrm{pN\,nm\,rad}^{-1}$. We performed a weighted least-squares fit of the plasmid torque τ_0 to the kinetic cleavage data, with cost function

$$C = \sum_i w_i\,C_i = \sum_i w_i \sum_t \left(\left(\hat{f}_{\mathrm{clv},i}(t) - f_{\mathrm{clv},i}(t) \right)^2 - \mathrm{RSS}_{\mathrm{clv},i} \right). \tag{4.2}$$

Here, C_i is the partial fit cost to target i, which scores the fit between the predicted cleaved fraction $\hat{f}_{\mathrm{clv},i}(t)$ and the measured cleaved fraction $f_{\mathrm{clv},i}(t)$. The value C_i compares this score to $\mathrm{RSS}_{\mathrm{clv},i}$, the residual sum of squares of individual exponential fit curves $y_{\mathrm{clv},i} = 1 - \exp(-k_{\mathrm{clv},i} \cdot t)$ to each target i. The target weights w_i were defined such that the perfect target ($w_i = 1$), single-mismatches ($w_i = 1/60$) and double-mismatches ($w_i = 1/851$) all contributed equally to the fit cost. After obtaining a fit value for τ_0, we expressed the corresponding cleavage predictions in terms of the effective cleavage rates $k_{\mathrm{clv},i}$ obtained with an exponential fit to the predicted cleavage curves.

4.5. Author contributions

Conceptualization: I.J., H.S.O., M.D., S.K.J. Methodology: I.J., H.S.O., M.D., S.K.J. Software: H.S.O., M.V., M.D., S.K.J. Validation: I.J., H.S.O., U.B. Formal analysis: I.J., H.S.O., M.V., U.B. Investigation: I.J., H.S.O., M.V., U.B. Resources: S.K.J., M.D. Data curation: I.J., H.S.O. Writing: I.J., H.S.O., M.D., S.K.J. Editing: I.J., H.S.O., M.V., M.D., S.K.J. Supervision: I.J., S.K.J., M.D. Project administration: M.D., S.K.J. Funding acquisition: S.K.J., M.D.

4.6. Supplementary Figures

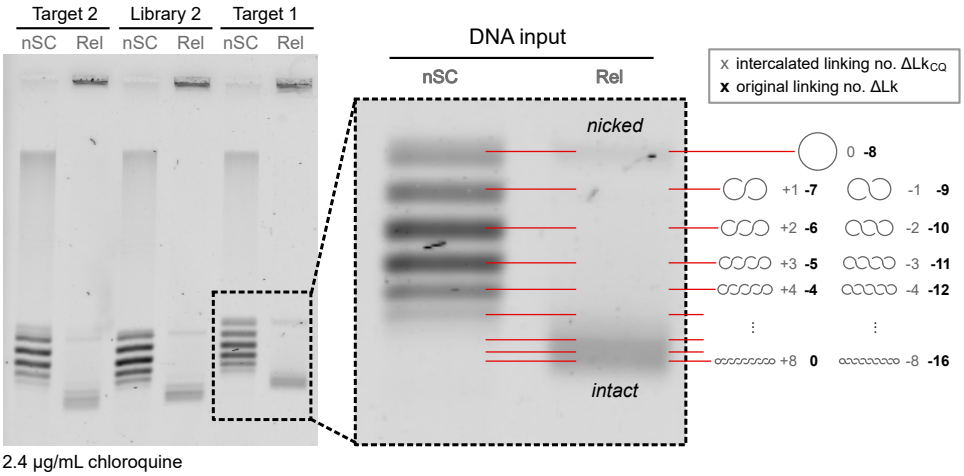


Figure 4.S1: Gel-based superhelical density determination. Illustration of supercoiling quantification method for Target-1 plasmids. The supercoiled and relaxed plasmid DNA inputs were ran on an agarose gel mixed with 2.4 µg/mL of chloroquine in TBE buffer with the same concentration of chloroquine (**Methods 4.4**). Band positions in the gel arise from the linking number differences at the experimental chloroquine level, $\Delta\mathrm{Lk}_{\mathrm{CQ}} = \Delta\mathrm{Lk} + f([\mathrm{CQ}])$. Here, the original linking number difference $\Delta\mathrm{Lk} = \mathrm{Lk} - L/h_0$ applies to a plasmid of length $L = 1.7\,\mathrm{kbp}$ and B-DNA pitch $h_0 = 10.5\,\mathrm{bp}$ in absence of chloroquine. The function $f([\mathrm{CQ}])$ captures how chloroquine increases DNA pitch beyond h_0. In the lane with the originally relaxed plasmids, the center of the intact-plasmid bands (labeled "intact") revealed $\Delta\mathrm{Lk} = 0$, and the nicked plasmids in the population (labeled "nicked") showed $\Delta\mathrm{Lk}_{\mathrm{CQ}} = 0$. Based on the total band count, we estimated $f([\mathrm{CQ}]) \approx 8$. With this value, each band in the supercoiled (nSC) population corresponded to two possible $\Delta\mathrm{Lk}$ values, as 1D electrophoresis cannot discern positive and negative supercoiling [46]. In line with previously measured $\sigma = -0.032 \pm 0.005$ for 2–50 kb plasmids in *E. coli* [77], we chose the lower value for $\Delta\mathrm{Lk}$. Next, we obtained the band intensity profiles (see **Methods 4.4**) and estimated the distribution mean and standard deviation by maximum likelihood estimation. This approach resulted in $\Delta\mathrm{Lk} = -5.29 \pm 1.15$ for Target 2, $\Delta\mathrm{Lk} = -4.77 \pm 1.14$ for Library 2, and $\Delta\mathrm{Lk} = -5.28 \pm 1.31$ for Target 1. A weighted average of these three results yielded a superhelical density of $\sigma = \Delta\mathrm{Lk}/(L/h_0) = -0.033 \pm 0.007$.

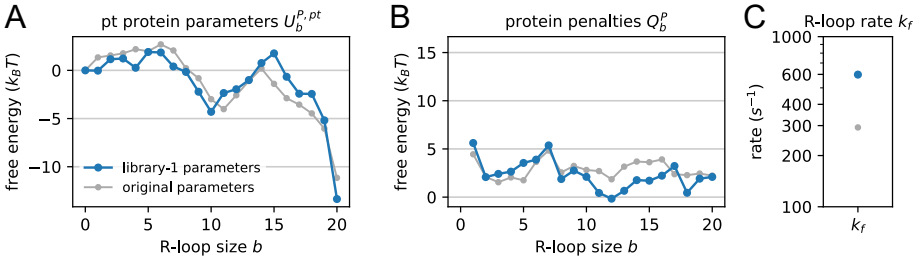


Figure 4.S2: Custom CRISPRzip parameter set. Library-1-specific CRISPRzip parameter values obtained for this study (blue; **Methods 4.4**), alongside original parameter values (gray). The parameter sets comprise 20 perfect-target protein parameters $U_b^{\mathrm{P,pt}}$ (**A**), 20 protein penalties Q_b^{P} (**B**) and 1 R-loop extension rate k_{f} (**C**).

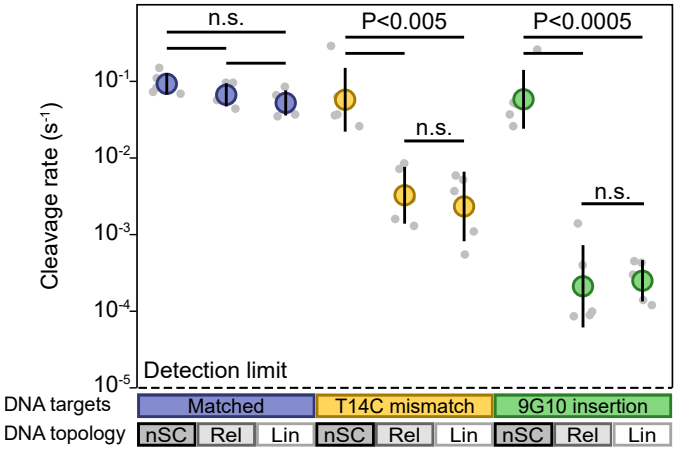


Figure 4.S3: Gel-obtained cleavage rates of supercoiled vs relaxed on- and off-targets. Data is present as mean±SD, N=5 in each group. One-Way ANOVA with Tukey's post-hoc test.

4.7. Supplementary Data

Name	RNA sequence (5' to 3')
Target 1 (R11)	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUC CGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU GACGCAUAAAGAUGAGACGC
sgRNA 2	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUC CGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU GUGAUAAGUGGAAUGCCAUG

Table 4.S1: sgRNAs used in the study. Target sequence in red, scaffold in black.

Name	DNA sequence (5' to 3')
Pr7.26_Target 2 library forward	GCTCTTCCGATCTAACCGCCGAATAACAGAGT
Pr7.27_Target 2 library reverse	TCTTCCGATCTAAGAACGCCTCGCACACT
Pr7.29_Vector 2 forward	GGCGTTCTTAGATCGGAAGAGCACACGTC
Pr7.28_Vector 2 reverse	GGCGGTTAGATCGGAAGAGCGTCGTGT
Pr6.29_Target 1 library forward	ACACGACGCTCTTCCGATCTATAACTAATTGAGCTGAACGCA C
Pr6.30_Target 1 library reverse	GACGTGTGCTCTTCCGATCTCTGAATAGTCGTGTAGTTGTGC T
Pr6.32_Vector 1 forward	AGATCGGAAGAGCACACGTC
Pr6.31_Vector 1 reverse	AGATCGGAAGAGCGTCGTGT
Pl0.63_Vector	AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTTGTTCTT TTGCACTACCGTCAGGTAATCTCGTATGCCGTCTTCTGCTTG CCCGGGTTTAAACACATGTGTTTTTCCATAGGCTCCGCCCCC CTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGG CGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCC TGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGC TTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTG GCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTG TAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCC CGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTC TTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCA GCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTA...

Name	DNA sequence (5' to 3')
Pl0.63_Vector (continued)	...GGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACG GCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGA AGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCG GCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCA AGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGAT CCTTTGATCTTTTCTACGTTACGCCCCGCCCTGCCACTCATC GCAATACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAA GCCGTCACAGACGGCATGATGAACCTGAATCGCCAGCGGCA TCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGA AAACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAA TCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAA AAACATATTCTCAATAAACCCTTTAGGGAAATAGGCCAGGTT TTCACCGTAACACGCCACATCTTGCGAATATATGTGTAGAAA CTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAA ACGTTTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAA CACTATCCCATATCACCAGCTCACCGTCTTTCATTGCCATAC GGAATTCCGGATGAGCATTCATCAGGCGGGCAAGAATGTGA ATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCT TTAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTTATAGG TACATTGAGCAACTGACTGAAATGCCTCAAAATGTTCTTTAC GATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTT TTTTCTCCATACTCTTCCTTTTTCAATATTATTGAAGCATTTAT CAGGGTTATTGTCTCATGAGCGGATACATATGGGTTACGCGT CGCGAGGCCATATGGGTTAACCCATGGCCAAGCTTGCATGC CTGCAGGTCGACGGGCCCGGGAATGATACGGCGACCACCG AGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Table 4.S2: Primers and plasmid backbone used. From Metabion or Azenta.

Name	DNA sequence (5’ to 3’)
Pr.2.21_Perfect target 1	ATAACTAATTGAGCTGAACGCACCAACGTGGTACTCTGTGCA GATTTTAGACGCATAAAGATGAGACGCTGGTGATCGTCTTAG AGACAAGCACAGTCTCTCCACGAAGCATTACTCTAGCACAAC TACACGACTATTCAG
Pr6.77_T14C	ATAACTAATTGAGCTGAACGCACGCGCTTGGTATCTTGCCCA GATATTAGACGCACAAAGATGAGACGCTGGTGATCGTCTTA GAGACAAGCACAGTCTCTCCTCTTCTATTGATGTCAGCACAA CTACACGACTATTCAG
Pr6.79_9G10	ATAACTAATTGAGCTGAACGCACGTGCCGTACCTGTTGCTCA GATTTTAGGACGCATAAAGGATGAGACGCTGGTGATCGTCTT AGAGACAAGCACAGTCTTTGACGCGAACGAGAGCAGCACAA CTACACGACTATTCAG
Pr7.09_Δ3	ATAACTAATTGAGCTGAACGCACCGGACTAGGAGAGAGAGC AGATTTTAGACGCATAAAGATGAGAGCTGGTGATCGTCTTAG AGACAAGCACAGTCTCTGCGGATCTTGACTGCATTAGCACAA CTACACGACTATTCAG
Pr7.32_3A4	ATAACTAATTGAGCTGAACGCACTCAAGACAAGACGCTGGC AGATTATTAGACGCATAAAGATGAGAACGCTGGTGATCGTCT TAGAGACAAGCACAGTCTGAATCCATCGAGAACTGAGCACA ACTACACGACTATTCAG
Pr8.54_SDM C3G forward	GACGCATAAAGATGAGAGGCTGGTGATCGTC
Pr8.55_SDM C3G reverse	GACGATCACCAGCCTCTCATCTTTATGCGTC
Pr7.50_Cy5 labeled forward	/5ATTO647NN/ACACGACGCTCTTCCGATCTATAACTAATTGA GCTGAACGCAC
Pr8.27_Cy3 labeled reverse	/Cy3/GACGTGTGCTCTTCCGATCTCTGAATAGTCGTGTAGTT GTGCT
Pl1.70_Perfect target 1	ACGCGTCGCGAGGCCATATGGGTTAACCCATGGCCAAGCTT GCATGCCTGCAGGTCGACGGGCCCGGGAATGATACGGCGA CCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCTATAACTAATTGAGCTGAACGCACCAACGTGGTACTCTGT GCAGATTTTAGACGCATAAAGATGAGACGCTGGTGATCGTCT TAGAGACAAGCACAGTCTCTCCACGAAGCATTACTCTAGCAC AACTACACGACTATTCAGAGATCGGAAGAGCACACGTCTGA ACTCCAGTCACTTGTTCTTTTGCACTACCGTCAGGTAATCTC GTATGCCGTCTTCTGCTTGCCCGGGTTTAAACACATGTGTTT TTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCG ACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAA GATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCT CCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTT TCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCT GTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGC...

Name	DNA sequence (5' to 3')
PI1.70_Perfect target 1 (continued)	...TGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCT GCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA GACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGG ATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTT GAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTAT TTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAA GAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGT AGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGA AAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGTTAC GCCCCGCCCTGCCACTCATCGCAATACTGTTGTAATTCATTA AGCATTCTGCCGACATGGAAGCCGTCACAGACGGCATGATG AACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCG TATAATATTTGCCCATGGTGAAAACGGGGGCGAAGAAGTTGT CCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCC AGGGATTGGCTGAGACGAAAAACATATTCTCAATAAACCCTT TAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTT GCGAATATATGTGTAGAAACTGCCGGAAATCGTCGTGGTATT CACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAAA CGGTGTAACAAGGGTGAACACTATCCCATATCACCAGCTCA CCGTCTTTCATTGCCATACGGAATTCCGGATGAGCATTCATC AGGCGGGCAAGAATGTGAATAAAGGCCGGATAAAACTTGTG CTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATATCCAGC TGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAAT GCCTCAAAATGTTCTTTACGATGCCATTGGGATATATCAACG GTGGTATATCCAGTGATTTTTTTCTCCATACTCTTCCTTTTTC AATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCG GATACATATGGGTT

Table 4.S3: Single targets construction primers.

References

1. Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607. doi:10.1038/nature09886 (2011).
2. Nishimasu, H. et al. Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* **156,** 935–949. doi:10.1016/j.cell.2014.02.001 (2014).
3. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821. doi:10.1126/science.1225829 (2012).
4. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67. doi:10.1038/nature13011 (2014).
5. Jiang, F. & Doudna, J. A. CRISPR–Cas9 Structures and Mechanisms. *Annual Review of Biophysics* **46,** 505–529. doi:10.1146/annurev-biophys-062215-010822 (2017).

6. Zeng, Y. et al. The initiation, propagation and dynamics of CRISPR-SpyCas9 R-loop complex. *Nucleic Acids Research* **46,** 350–361. doi:10.1093/nar/gkx1117 (2018).
7. Kerwash, E. & Johnston, J. D. Casgevy: Innovative Medicinal Products Require Innovative Approaches to Regulatory Assessment. *Pharmaceutics* **16,** 906. doi:10.3390/pharmaceutics16070906 (2024).
8. Musunuru, K. et al. Patient-Specific In Vivo Gene Editing to Treat a Rare Genetic Disease. *New England Journal of Medicine* **392,** 2235–2243. doi:10.1056/NEJMoa2504747 (2025).
9. Singh, A. et al. Revolutionary breakthrough: FDA approves CASGEVY, the first CRISPR/Cas9 gene therapy for sickle cell disease. *Annals of Medicine and Surgery* **86,** 4555. doi:10.1097/MS9.0000000000002146 (2024).
10. Cullot, G. et al. CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nature Communications* **10,** 1–14. doi:10.1038/s41467-019-09006-2 (2019).
11. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology* **31,** 822–827. doi:10.1038/nbt.2623 (2013).
12. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* **31,** 827–832. doi:10.1038/nbt.2647 (2013).
13. Zhang, Y. et al. Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. *Scientific Reports* **4,** 5405. doi:10.1038/srep05405 (2014).
14. Jones, S. K. et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nature Biotechnology* **39.** ISBN: 4158702006465, 84–93. doi:10.1038/s41587-020-0646-5 (2021).
15. Lin, Y. et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research* **42,** 7473–7485. doi:10.1093/nar/gku402 (2014).
16. Newton, M. D. et al. Negative DNA supercoiling induces genome-wide Cas9 off-target activity. *Molecular Cell* **83,** 3533–3545.e5. doi:10.1016/j.molcel.2023.09.008 (2023).
17. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature Biotechnology* **31,** 839–843. doi:10.1038/nbt.2673 (2013).
18. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology* **33,** 187–197. doi:10.1038/nbt.3117 (2015).
19. Wienert, B. et al. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science* **364,** 286–289. doi:10.1126/science.aav9023 (2019).

20. Zou, R. S. et al. Improving the sensitivity of in vivo CRISPR off-target detection with DISCOVER-Seq+. *Nature Methods* **20,** 706–713. doi:10.1038/s41592-023-01840-z (2023).

21. Mali, P. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnology* **31,** 833–838. doi:10.1038/nbt.2675 (2013).

22. Boyle, E. A. et al. Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Science Advances* **7.** doi:10.1126/sciadv.abe5496 (2021).

23. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34,** 184–191. doi:10.1038/nbt.3437 (2016).

24. Listgarten, J. et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering* **2,** 38–47. doi:10.1038/s41551-017-0178-6 (2018).

25. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods* **12,** 982–988. doi:10.1038/nmeth.3543 (2015).

26. Abadi, S., Yan, W. X., Amar, D. & Mayrose, I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLOS Computational Biology* **13,** e1005807. doi:10.1371/journal.pcbi.1005807 (2017).

27. Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30,** 1473–1475. doi:10.1093/bioinformatics/btu048 (2014).

28. Eslami-Mossallam, B. et al. A kinetic model predicts SpCas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity. *Nature Communications* **13.** doi:10.1038/s41467-022-28994-2 (2022).

29. Lin, J. & Wong, K.-C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics* **34,** i656–i663. doi:10.1093/bioinformatics/bty554 (2018).

30. 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 Computational Biology* **15,** e1007480. doi:10.1371/journal.pcbi.1007480 (2019).

31. Luo, Y., Chen, Y., Xie, H., Zhu, W. & Zhang, G. Interpretable CRISPR/Cas9 off-target activities with mismatches and indels prediction using BERT. *Computers in Biology and Medicine* **169,** 107932. doi:10.1016/j.compbiomed.2024.107932 (2024).

32. Peng, H., Zheng, Y., Zhao, Z., Liu, T. & Li, J. Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions. *Bioinformatics* **34,** i757–i765. doi:10.1093/bioinformatics/bty558 (2018).

33. Ivanov, I. E. et al. Cas9 interrogates DNA in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 5853–5860. doi:10.1073/pnas.1913445117 (2020).

34. Newton, M. D. et al. DNA stretching induces Cas9 off-target activity. *Nature Structural and Molecular Biology* **26,** 185–192. doi:10.1038/s41594-019-0188-z (2019).

35. Janissen, R., Barth, R., Polinder, M., van der Torre, J. & Dekker, C. Single-molecule visualization of twin-supercoiled domains generated during transcription. *Nucleic Acids Research* **52,** 1677–1687. doi:10.1093/nar/gkad1181 (2024).

36. Fogg, J. M., Judge, A. K., Stricker, E., Chan, H. L. & Zechiedrich, L. Supercoiling and looping promote DNA base accessibility and coordination among distant sites. *Nature Communications* **12,** 5683. doi:10.1038/s41467-021-25936-2 (2021).

37. Zeida, A., Machado, M. R., Dans, P. D. & Pantano, S. Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations. *Physical Review E* **86,** 021903. doi:10.1103/PhysRevE.86.021903 (2012).

38. Murugan, K., Suresh, S. K., Seetharam, A. S., Severin, A. J. & Sashital, D. G. Systematic in vitro specificity profiling reveals nicking defects in natural and engineered CRISPR–Cas9 variants. *Nucleic Acids Research* **49,** 4037–4053. doi:10.1093/nar/gkab163 (2021).

39. Yan, W. X. et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications* **8,** 15058. doi:10.1038/ncomms15058 (2017).

40. Schmid-Burgk, J. L. et al. Highly Parallel Profiling of Cas9 Variant Specificity. *Molecular Cell* **0.** doi:10.1016/j.molcel.2020.02.023 (2020).

41. Hawkins, J. A., Jones, S. K., Finkelstein, I. J. & Press, W. H. Indel-correcting DNA barcodes for high-throughput sequencing. *Proceedings of the National Academy of Sciences* **115,** E6217–E6226. doi:10.1073/pnas.1802640115 (2018).

42. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573. doi:10.1038/nature13579 (2014).

43. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences* **114,** 5461–5466. doi:10.1073/pnas.1700557114 (2017).

44. Bravo, J. P. K. et al. Structural basis for mismatch surveillance by CRISPR–Cas9. *Nature.* doi:10.1038/S41586-022-04470-1 (2022).

45. Gong, S., Yu, H. H., Johnson, K. A. & Taylor, D. W. DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell Reports* **22,** 359–371. doi:10.1016/J.CELREP.2017.12.041 (2018).

46. Martínez-García, B. et al. in *Molecular Motors: Methods and Protocols* (ed Lavelle, C.) 291–300 (Springer, New York, NY, 2018). ISBN: 978-1-4939-8556-2. doi:10.1007/978-1-4939-8556-2_15.

47. Pacesa, M. et al. Structural basis for Cas9 off-target activity. *Cell* **185,** 4067–4081.e21. doi:10.1016/j.cell.2022.09.026 (2022).

48. Yang, M. et al. The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET. *Cell Reports* **22,** 372–382. doi:10.1016/j.celrep.2017.12.048 (2018).

49. Pacesa, M. et al. R-loop formation and conformational activation mechanisms of Cas9. *Nature* **609.** Number: 7925, 191–196. doi:10.1038/s41586-022-05114-0 (2022).

50. Hua Fu, B. X. et al. Target-dependent nickase activities of CRISPR-Cas nucleases Cpf1 and Cas9. *Nature microbiology* **4,** 888–897. doi:10.1038/s41564-019-0382-0 (2019).

51. Fan, R. et al. Shortening the sgRNA-DNA interface enables SpCas9 and eSpCas9(1.1) to nick the target DNA strand. *Science China Life Sciences* **63,** 1619–1630. doi:10.1007/s11427-020-1722-0 (2020).

52. Lazzarotto, C. R. et al. CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity. *Nature Biotechnology* **38,** 1317–1327. doi:10.1038/s41587-020-0555-7 (2020).

53. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nature Methods* **14,** 607–614. doi:10.1038/nmeth.4278 (2017).

54. Wu, X. et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology* **32.** Number: 7, 670–676. doi:10.1038/nbt.2889 (2014).

55. Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G. F. & Chin, L. Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. *Cell Reports* **14,** 1555–1566. doi:10.1016/j.celrep.2016.01.019 (2016).

56. Van Overbeek, M. et al. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell* **63,** 633–646. doi:10.1016/j.molcel.2016.06.037 (2016).

57. Hinz, J. M., Laughery, M. F. & Wyrick, J. J. Nucleosomes Inhibit Cas9 Endonuclease Activity in Vitro. *Biochemistry* **54,** 7063–7066. doi:10.1021/acs.biochem.5b01108 (2015).

58. Horlbeck, M. A. et al. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife* **5** (ed Adelman, K.) e12677. doi:10.7554/eLife.12677 (2016).

59. Isaac, R. S. et al. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *eLife* **5,** e13450. doi:10.7554/eLife.13450 (2016).

60. Yarrington, R. M., Verma, S., Schwartz, S., Trautman, J. K. & Carroll, D. Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proceedings of the National Academy of Sciences* **115,** 9351–9358. doi:10.1073/pnas.1810062115 (2018).

61. Knight, S. C. et al. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350,** 823–826. doi:10.1126/science.aac6572 (2015).

62. Wang, A. S. et al. The Histone Chaperone FACT Induces Cas9 Multi-turnover Behavior and Modifies Genome Manipulation in Human Cells. *Molecular Cell* **79,** 221–233.e5. doi:10.1016/j.molcel.2020.06.014 (2020).

63. Kim, D. et al. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nature Biotechnology* **34,** 863–868. doi:10.1038/nbt.3609 (2016).

64. Strohkendl, I., Saifuddin, F. A., Rybarski, J. R., Finkelstein, I. J. & Russell, R. Kinetic Basis for DNA Target Specificity of CRISPR-Cas12a. *Molecular Cell* **71,** 816–824.e3. doi:10.1016/j.molcel.2018.06.043 (2018).

65. Aris, K. D. P. et al. Dynamic basis of supercoiling-dependent DNA interrogation by Cas12a via R-loop intermediates. *Nature Communications* **16,** 2939. doi:10.1038/s41467-025-57703-y (2025).

66. Parameshwaran, H. P. et al. The bridge helix of Cas12a imparts selectivity in cis-DNA cleavage and regulates trans-DNA cleavage. *FEBS Letters* **595,** 892–912. doi:10.1002/1873-3468.14051 (2021).

67. Van Aelst, K., Martínez-Santiago, C., Cross, S. & Szczelkun, M. The Effect of DNA Topology on Observed Rates of R-Loop Formation and DNA Strand Cleavage by CRISPR Cas12a. *Genes* **10,** 169. doi:10.3390/genes10020169 (2019).

68. Aldag, P. et al. Dynamic interplay between target search and recognition for a Type I CRISPR-Cas system. *Nature Communications* **14.** Number: 1, 3654. doi:10.1038/s41467-023-38790-1 (2023).

69. Rutkauskas, M. et al. Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Reports* **10,** 1534–1543. doi:10.1016/j.celrep.2015.01.067 (2015).

70. Westra, E. R. & Levin, B. R. It is unclear how important CRISPR-Cas systems are for protecting natural populations of bacteria against infections by mobile genetic elements. *Proceedings of the National Academy of Sciences* **117,** 27777–27785. doi:10.1073/pnas.1915966117 (2020).

71. Shi, Y.-J. et al. DNA topology regulates PAM-Cas9 interaction and DNA unwinding to enable near-PAMless cleavage by thermophilic Cas9. *Molecular Cell* **82,** 4160–4175.e6. doi:10.1016/j.molcel.2022.09.032 (2022).

72. Huo, Y. et al. Structural and biochemical insights into the mechanism of the Gabija bacterial immunity system. *Nature Communications* **15,** 836. doi:10.1038/s41467-024-45173-7 (2024).

73. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576,** 149–157. doi:10.1038/s41586-019-1711-4 (2019).

74. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353,** aaf8729. doi:10.1126/science.aaf8729 (2016).

75. Hammer, Ø., Harper, D. A. T. & Ryan, P. D. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica* **4,** 1–9 (2001).

76. Gao, X., Hong, Y., Ye, F., Inman, J. T. & Wang, M. D. Torsional Stiffness of Extended and Plectonemic DNA. *Physical Review Letters* **127,** 028101. doi:10.1103/PhysRevLett.127.028101 (2021).

77. Wang, J. C. Degree of superhelicity of covalently closed cyclic DNA's from *Escherichia coli*. *Journal of Molecular Biology* **43,** 263–272. doi:10.1016/0022-2836(69)90266-6 (1969).

5

From molecular Cas9 kinetics to genome-wide cleavage patterns in cells

CRISPR-Cas9 is widely used for genome editing, yet predicting its dynamics in vivo *remains challenging due to the great number of genomic sites that it can potentially bind and cleave. In this chapter, we assess how heterogeneous genome interactions influence the kinetics of target search and cleavage dynamics. Using CRISPRzip, a mechanistic R-loop model trained* in vitro*, we make general predictions how Cas9 interrogates and cleaves genomic targets, depending on DNA supercoiling state. Our predictions show that one can often neglect the influence of genome sequence on the expected cleavage dynamics, whether due to target search delay, competitive cleavage dynamics, or background cleavage activity. We introduce the cleavage capacity as an abstract quantity that unifies editing outcomes across delivery strategies and experimental implementations. Overall, our work demonstrates how molecular-scale measurements of CRISPR nucleases can be translated to genome-wide predictions, providing a biophysically grounded framework for CRISPR applications.*

5.1. Introduction

Much research has been conducted to map out Cas9's on- and off-target cleavage activity as a function of gRNA choice, as summarized in **Chapter 1**. Most of these efforts adopt a top-down strategy, studying the cleavage patterns that emerge from different gRNAs in specific cellular contexts [1–4]. This strategy has resulted in a wide range of activity prediction algorithms that are valuable to gRNA design, however, these models generally lack predictive power outside the context of the original experiments [1, 5–8].

In contrast, *in vitro* and single-molecule studies pursue a bottom-up strategy, isolating the molecular mechanisms of Cas9 under controlled conditions. Thanks to these efforts, we understand in detail how Cas9 interacts with individual DNA targets, in terms of target search *in vivo* [9–12] and *in vitro* [13–15], target interrogation [16–20], and effects of DNA state [21, 22]. Meanwhile, high-throughput assays [23–26] have greatly increased the number of sequences for which high-quality kinetic data is available.

Using these molecular insights, biophysicists have constructed models that reveal the universal principles of RNA-guided target search, and as such do generalize to different experimental conditions. Some biophysical models assume a rapid binding equilibrium on the genome, such that cleavage of a site directly follows from the free energy cost for Cas9 to bind that site [27–29]. However, *in vivo* experiments show that DNA dissociation is not fast enough to guarantee such an equilibrium [10–12] and it has been demonstrated that discrimination between DNA targets depends more strongly on the energetic barrier to specific target binding than on the final binding stability [30]. These observations indicate a need for kinetic models of target search [31]. Efforts to construct kinetic models include those by Shvets and Kolomeisky, using a first-passage time approach [32], Lu and Pigoletti, applying the facilitated diffusion framework [33, 34], and Zhang et al., employing the King-Altman method [35]. However, a special and defining feature of Cas9 reaction kinetics is how it heterogeneously interacts with the many and diverse binding sites in a genome, and the existing models give little insight into this aspect of its dynamics. The Lu and Pigoletti model considers variable PAM binding but no R-loop formation beyond the intended target [34], the Shvets and Kolomeisky model assumes a uniform off-target cleavage rate [32], and the Zhang et al. model uses neural networks to predict off-target heterogeneity and to generalize from *in vitro* to *in vivo*, which lacks biophysical or mechanistic support [35].

This work presents a kinetic model of Cas9 activity *in vivo* that explicitly considers Cas9's heterogeneous interactions with the whole genome. We use CRISPRzip, a mechanistic model of Cas9 target recognition (**Chapter 3**), to predict heterogeneous R-loop kinetics, and we translate its *in vitro* predictions to an *in vivo* context (**Chapter 2**). Crucially, CRISPRzip explicitly considers how DNA torque or superhelical density affects target recognition, which we here show to be a major determinant of its *in vivo* kinetics.

First, we investigate if the contents of the genome modulate the cleavage dynamics of particular DNA targets. One potential effect due to genome contents is target search delay due to partially matching DNA sites that trap Cas9. Such slowdown has been experimentally demonstrated for synthetic genomes [36], but we show that this phenomenon does not likely occur in wild-type genomes. Another hypothetical effect occurs when the genome contains many potential targets, each of which Cas9 can cleave fast and remain stably bound to, giving rise to complex competition-like DSB

dynamics. We predict the magnitude of this effect and argue it is irrelevant compared to Cas9 degradation rates in human cells.

Second, having arrived at a simple model that ignores the genetic background, we predict DSB dynamics in human cells due to either exogenous Cas9 expression or nuclear Cas9 delivery. The different editing outcomes in these settings can be unified with a single quantity, the "cleavage capacity" of an assay. Finally, we explore which DNA sites contribute most to the total off-target cleavage: the many weak off-targets, or the few strong off-targets in a genome. There is a balance between off-target presence and strength that shifts with time and depends on assay conditions, and this balance generally favors the few strong off-targets in a genome. Accordingly, we expect no significant background cleavage, i.e., many rare DSBs that would go undetected in cleavage assays.

In summary, this work shows that genome contents do not affect expected cleavage dynamics and do not give rise to many, hardly predictable background DSBs. In addition, it introduces the cleavage capacity as an abstract quantity to predict editing outcomes independent of experimental implementation. Together, these findings support the use of *in vitro* data and models to make direct predictions about *in vivo* Cas9 applications, and help to connect our molecular understanding of CRISPR-Cas nucleases to the complex reality of the genome.

5.2. Results

5.2.1. Genomes typically contain too few partial matches to delay target search

The time for Cas9 to locate its target in a cell depends on the surrounding genome. Intuitively, larger genomes take longer to explore, but it is unclear if the genome sequence affects target search time as well. It has been proposed that partially matching non-target sites can temporarily trap Cas9 complexes and hence delay their arrival at the target site [36, 37]. We explore the collective trapping capacity of these non-target sites according to a kinetic target search model. This non-target trapping effect is separate from the effects of cleavage on Cas9 availability and target-search time; here, we study trapping for a binding-only Cas9 system, and in **Section 5.2.3**, we investigate cleavage effects.

We represent the process of target search by Cas9 in a Markov model and solve the first-passage time (FPT) problem of a single Cas9 complex arriving at one particular PAM site i (**Figure 5.1A**). A single free Cas9 ribonucleoprotein (RNP) binds any PAM site on the genome at a rate k_{bnd} and unbinds it at a rate k_{unb}. In a genome containing N PAM sites, a free RNP complex binds an individual PAM site at a rate k_{bnd}/N. Because we currently study non-target trapping as an isolated binding-only effect, we assume all sites are non-targets that cannot be cleaved by Cas9. If Cas9 forms a partial R-loop on a non-target, it does so on the first ~8 bp, comprising the so-called "seed". The processes of R-loop formation and subsequent rejection of the seed take place at rates $k_{\text{seed},i}$ and $k_{\text{reject},i}$. The large majority of DNA sites have too little complementarity with the gRNA to allow for partial R-loop formation, implying that their $k_{\text{seed},i}$ is negligable. We derive the

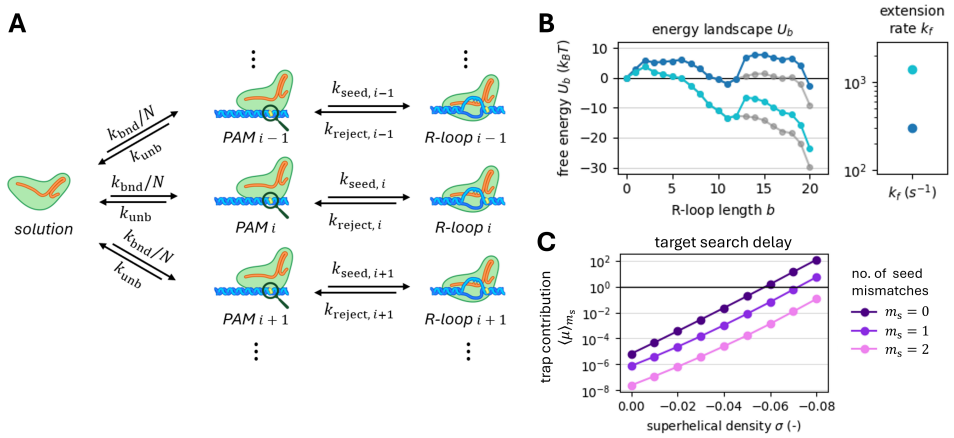


Figure 5.1: Genomes typically contain too few partial matches to delay target search. A Kinetic model of target search by a single Cas9 complex in presence of partially matching non-target sites i that can delay arrival at the target PAM. **B** Free-energy landscape predicted by CRISPRzip for a site with a mismatching bp located at position 12 (blue) compared to the perfect-target landscape (gray). At a superhelical density of $\sigma = -0.05$, the energy landscape and the R-loop extension rate change (cyan), promoting R-loop formation. **C** Contributions $\langle\mu\rangle_{m_s}$ to the genome's total trap capacity $\langle\mu\rangle$ of all patterns with $m_s = 0, 1, 2$ seed mismatches (purple, violet, pink). The perfect-seed contribution $\langle\mu\rangle_0$ makes up the largest contribution to the total trap capacity of the full genome, and accordingly dominates the target search delay.

average FPT of a single RNP at a particular PAM i (**Suppl. Note 5.S1**) and find

$$\langle T^1_{\mathrm{PAM}}\rangle = N\left(\frac{1}{k_{\mathrm{bnd}}} + \frac{1}{k_{\mathrm{unb}}} + \frac{1}{k_{\mathrm{unb}}}\frac{1}{N}\sum_{i=1}^{N}\frac{k_{\mathrm{seed},i}}{k_{\mathrm{reject},i}}\right). \tag{5.1}$$

This expression is identical to the expected target search time in absence of traps (**Chapter 2**, **Equation 2.5** with $\langle T^1_{\mathrm{PAM}}\rangle = 1/k_{\mathrm{PAM}}$ and $C = 1$) apart from the presence of the last term within parentheses. Only when this last term becomes significantly larger than the first two, target search is delayed due to trapping by non-targets. We define the fraction between the seed-binding and -rejecting rates as the "trapping capacity" μ_i of non-target i. It can be associated with the effective stability of the non-target's partial R-loop state $\Delta U_{\mathrm{seed},i}$ through the Boltzmann factor

$$\mu_i = \frac{k_{\mathrm{seed},i}}{k_{\mathrm{reject},i}} = \exp\left(-\frac{\Delta U_{\mathrm{seed},i}}{k_B T}\right). \tag{5.2}$$

Using this definition, we rewrite **Equation 5.1** as an explicit condition for trapping to cause target search delay, yielding

$$\langle\mu\rangle = \frac{1}{N}\sum_{i=1}^{N}\mu_i \gtrsim 1 + \frac{k_{\mathrm{unb}}}{k_{\mathrm{bnd}}}. \tag{5.3}$$

This condition shows that delay follows from the average trapping capacity $\langle\mu\rangle$ of all PAM sites. Hence, while the total search time $\langle T^1_{\text{PAM}}\rangle$ scales with genome length N, the relative delay due to trapping does not. Although longer genomes contain more non-targets that collectively trap Cas9 more strongly, they equally contain more sites on which Cas9 only interacts with the PAM, so their average trapping capacity $\langle\mu\rangle$ remains the same as for shorter genomes.

We quantify the total trapping strength of a genome with the mechanistic model CRISPRzip (**Chapter 3**). CRISPRzip predicts the target interrogation dynamics of Cas9 based on the free-energy landscapes of R-loop formation at base pair resolution (**Figure 5.1B**). It explicitly accounts for DNA superhelicity, which is a major determinant of a site's trapping capacity as negative supercoiling stabilizes partial R-loops. The original model is trained to consider the effects of nucleotide sequence on R-loop kinetics. Here however, we disregard such details and consider average predictions based on mismatch positions alone. Such sequence-average predictions allow us to estimate general R-loop dynamics for all possible gRNA-DNA sequences, and to consider many DNA sites while staying within computational limitations. We establish a sequence-average parameter set for CRISPRzip by training the adjusted model on the original dataset (**Methods 5.4**, **Figure 5.S1**).

Next, we calculate the expected presence of different non-targets on the genome. Rather than counting individual sites i, we consider mismatch patterns j, where each pattern is a particular sequence of 20 gRNA-DNA matching or mismatching base pairs. This definition is consistent with our sequence-average CRISPRzip model, which predicts R-loop dynamics based on such mismatch patterns. Each DNA site i is characterized by a mismatch pattern j, and each mismatch pattern j may be present in multiple sites i on the genome. To calculate the expected presence of pattern j, we consider a random genome in which all nucleotides are equally represented. Bernoulli statistics dictates the probability p_j that a random site has a mismatch pattern j with in total m mismatches: each nucleotide has an independent probability of $1/4$ to match and $3/4$ to mismatch the guide RNA, so

$$p_j = \left(\frac{3}{4}\right)^m \left(\frac{1}{4}\right)^{(20-m)}. \tag{5.4}$$

The expected presence of pattern j on a genome with N PAM sites is Np_j.

To estimate the total trapping capacity $\langle\mu\rangle$ of a genome, we first inspect the patterns j that trap Cas9 most strongly because they have few seed mismatches (in bps 1–8 of the target). With CRISPRzip, we estimate the individual trapping capacities μ_j of all 151,552 patterns j with ≤2 seed mismatches ($m_{\text{s}} \in \{0, 1, 2\}$), at variable supercoiling levels (**Methods 5.4**). Combining these values μ_j with the pattern probability p_j, we find the collective trapping contribution $\langle\mu\rangle_{m_{\text{s}}}$ of all patterns j with m_{s} seed mismatches: $\langle\mu\rangle_{m_{\text{s}}} = \sum_{j'} p_{j'}\mu_{j'}$, summing only over patterns j' with m_{s} seed mismatches. With this definition, the genome's total trapping capacity is split up according to $\langle\mu\rangle = \sum_{m_{\text{s}}}\langle\mu\rangle_{m_{\text{s}}}$.

Figure 5.1C shows that, at any considered superhelical density σ, the perfect-seed non-targets contribute most to the genome's total trap capacity, $\langle\mu\rangle_0 \gg \langle\mu\rangle_1, \langle\mu\rangle_2$. Although the non-targets with 1 or 2 seed mismatches outnumber the perfect-seed non-targets by a factor 8 and 28, the presence of seed mismatches reduces their trap

capacities μ_j more strongly. These findings suggest that a genome's total trap capacity $\langle\mu\rangle$ can be estimated from the expected perfect-seed non-targets only, $\langle\mu\rangle \approx \langle\mu\rangle_0$.

According to **Equation 5.3**, the total trap capacity $\langle\mu\rangle$ should be at least of order unity to delay target search. We observe that $\langle\mu\rangle_0$ meets this condition only at high levels of negative supercoiling $\sigma < -0.06$ (**Figure 5.1C**). We expect that genomes are typically not maintained at such uniformly high supercoiling levels. Experimental studies on wild type *Escherichia coli* (*E. coli*) measured genome-average values of σ between -0.03 and -0.05 [38–40] (**Chapter 4**). The DNA helicity of eukaryotic genomes is controlled in separate domains [41–43], and while some domains are maintained at unknown levels of negative supercoiling and could hypothetically trap Cas9, other domains are positively supercoiled and reduce the genome's average trapping capacity $\langle\mu\rangle$. In conclusion, we predict that non-targets do not significantly trap Cas9 and hence its target search time is unaffected by genome contents.

5.2.2. Modeling Cas9 trapping experiments in *E. coli*

We validate our predictions about delayed target search with experimental data in *E. coli* from Moreb et al. (2020) [36]. In these experiments, *E. coli* strains expressed Cas9 and gRNAs targeting different native genes in the host's own genome. Hence, successful target cleavage resulted in autoimmunity and was observable as reduced colony fitness. The bacteria were transformed with a plasmid containing 1–15 non-target sites that matched the first 8 bps of the intended target and partially matched the following 4 bps (**Figure 5.2A**). The presence of these plasmids at around 400 copies per cell could rescue colonies from autoimmunity depending on the total number of non-targets (**Figure 5.2B**).

We formulate a survival condition for these experiments based on **Equation 5.3**. This condition is expressed in terms of the effective stability of Cas9 after forming the 8-bp seed, $\Delta U_{\text{seed},j} = \log(k_{\text{reject},j}/k_{\text{seed},j})k_BT$, which on the selected targets corresponds to a partial 9–12 bp R-loop (**Figure 5.2A**). If N_T non-targets should delay target search by a factor λ in order to rescue a colony, they require a seed stability ΔU_{seed} of

$$\Delta U_{\text{seed}} \lesssim -\left(\ln\left(\frac{N_T}{N}\right) + \ln\left(1 + \frac{k_{\text{unb}}}{k_{\text{bnd}}}\right) + \ln(\lambda - 1)\right) k_BT. \tag{5.5}$$

The number of PAM sites N is estimated at 10^6 (two *E. coli* genome copies [10]), and we approximate $k_{\text{unb}} \approx k_{\text{bnd}}$ in line with single-molecule tracking experiments [11] and facilitated diffusion theory [44]. We cannot estimate the required λ value for colony rescue from the experiments, so we arbitrarily hypothesize that doubling the target search time ($\lambda = 2$) would save a colony from autoimmunity, and later assess if this requirement is reasonable. In this scenario, the observation that the *malG*-targeting colony survived at the presence of $N_T = 800 \pm 400$ non-targets (1–3 per plasmid, 400 plasmid copies) implies a stability of $\Delta U_{\text{seed}} = -8.0 \pm 0.5\, k_BT$. Similarly, the *pgaC* and *pheT*-targeting colonies would survive with $N_T = 5400 \pm 600$ non-targets with a stability $\Delta U_{\text{seed}} = -5.9 \pm 0.1\, k_BT$, and the *hisF* and *mqo*-targeting colonies would not survive with $N_T = 6000$ non-targets with a stability $\Delta U_{\text{seed}} > -5.8\, k_BT$ (**Figure 5.2C,D**).

We compare these seed stabilities with CRISPRzip to assess if the hypothetical

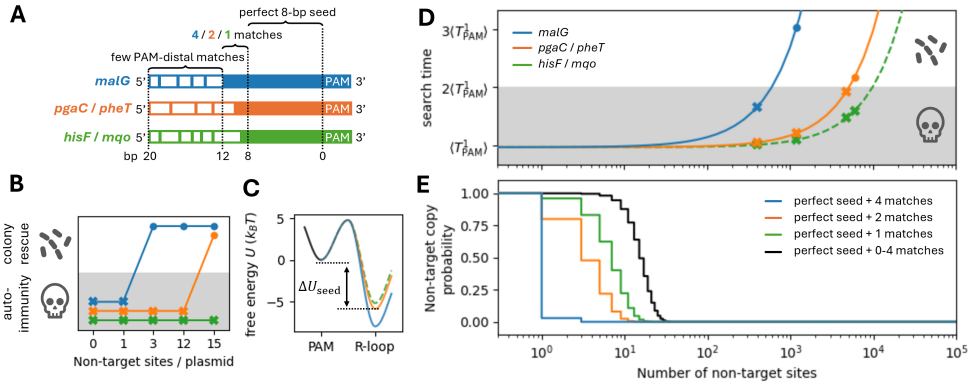


Figure 5.2: Modeling Cas9 trapping experiments in *E. coli*. **A** Targets in the host genome as selected in Moreb et al. (2020) [36]. **B** Colonies are rescued from autoimmunity depending on the copy number and trapping strength of non-targets. **C** Estimated R-loop stabilities ΔU_{seed} of non-target sites that slow down target search by a factor $\lambda = 2$. **D** Predicted time for a single RNP to find a single target as a function of non-target site abundance (**Equation 5.1**). Crosses and dots correspond to colony death and rescue as presented in panel B. **E** Poisson distribution describing non-target occurrence on two identical copies of a genome with 500,000 PAM sites. Lines indicate probability to find the indicated number of non-target sites or less.

rescue requirement $\lambda = 2$ is reasonable. Varying σ, we calculate the values $k_{\text{seed},i}/k_{\text{reject},i}$ for the mismatch patterns of *malG* and *pgaC* and find an optimal fit at $\sigma = -0.036$, with estimated $\Delta U_{\text{seed},malG} = -9.4\,k_BT$ and $\Delta U_{\text{seed},pgaC} = -5.7\,k_BT$. This superhelical density σ is in good agreement with experimental measurements of similarly large plasmids in *E. coli* ($\sigma = -0.033 \pm 0.007$ from **Chapter 4**, **Figure 4.S1**). This consistency suggests that a doubling of the target search time could indeed have saved colonies in these experiments conditions.

While the experiments prove that large non-target numbers can delay target search, it is prohibitively unlikely to find this many non-targets on a random genome. In an *E. coli* cell with 2 genomes, there is a 99.5% probability to have at most 30 PAM sites with a perfect seed, and most of these have too few matching bps beyond the seed to effectively trap Cas9 (**Figure 5.2E**). In conclusion, these experiments corroborate our numerical prediction that genomic non-target traps occur too rarely to substantially delay Cas9 target search under physiological conditions.

5.2.3. Human genome targets are cleaved independently, without competition

In cells, DSB profiles follow from the time evolution of free RNP availability (**Chapter 2**, **Equation 2.8**). The genome contents may influence these RNP levels, as Cas9 complexes remain stably bound to their targets after cleavage. For instance, single-molecule tracking in *E. coli* shows that catalytically dead Cas9 (dCas9) remains DNA-bound until a polymerase actively displaces it during replication [10]. When many

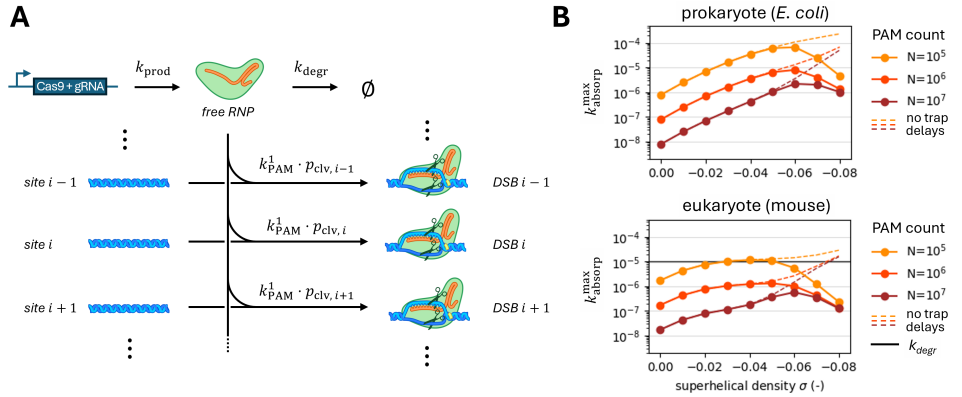


Figure 5.3: Human genome targets are cleaved independently, without competition for RNP. A Reaction scheme of an *in vivo* cleavage assay, with production and degradation. **B** Initial DSB-absorption rate coefficient $k^{\text{max}}_{\text{absorp}}$ according to **Equation 5.10**. Predictions for prokaryotes and eukaryotes based on estimates and measurements of PAM count N, and RNP binding and unbinding rates k_{bnd} and k_{unb} in *E. coli* and mice [9, 11]. Solid lines correspond to DSB absorption rates on genomes with variable number of available PAMs N. Dashed lines correspond to DSB absorption rates in absence of non-target traps, showing that target search delays occur at high negative supercoiling σ (**Figure 5.1C**). Black line in right panel indicates estimated degradation rate coefficient k_{degr} for eukaryotes.

RNPs get absorbed to DSBs and hence significantly deplete the pool of active RNPs, this results in complex, genome-dependent cleavage dynamics. Here, we investigate if this effect should be considered in Cas9 applications, in both prokaryotes (*E. coli*) and eukaryotes (mammals).

We model Cas9 dynamics *in vivo*, considering variable RNP levels and heterogeneous off-target cleavage (**Figure 5.3A**). The RNP copy number $C(t)$ evolves with time according to 1) the rate k_{prod} at which the cell produces RNP as a result of its gene expression; 2) the rate coefficient k_{degr} at which Cas9 is degraded by the cell; and 3) the total cleavage activity, as a result of which RNP gets absorbed to their DSBs. The total cleavage activity depends on the collective search dynamics of all RNPs and on the targets i present on the genome. When C Cas9 complexes carry out independent target search in parallel, a particular DNA site i gets PAM-bound at a rate that is linear in C, so we explicitly write $k_{\text{PAM}} = k^1_{\text{PAM}} \cdot C$. Here $k^1_{\text{PAM}} = 1/\langle T^1_{\text{PAM}} \rangle$ (**Equation 5.1**) is the effective rate coefficient of PAM binding, or equivalently, the effective PAM-binding rate due to single-RNP target search. We assume that target search is much slower than the time that Cas9 dwells on a PAM, such that the cleavage rate $k_{\text{clv},i}$ of site i can be expressed in the rare-binding limit (**Chapter 2**, **Equation 2.6**),

$$k_{\text{clv},i} = k^1_{\text{PAM}}\, p_{\text{clv},i} \cdot C \tag{5.6}$$

where $p_{\text{clv},i}$ is the probability for site i to get cleaved upon PAM binding.

To obtain the RNP dynamics and the DSB dynamics, we introduce a unitless

quantity Γ which is defined in terms of the anti-derivative of the RNP copy number, $\Gamma(t) = k^1_{\text{PAM}} \cdot \int_0^t dt'\, C(t')$. The value of $\Gamma(t)$ corresponds to the expected number of PAM-binding events at a particular site up to time t. As we argue later, Γ also quantifies the capacity of a Cas9 assay to induce DSBs, and as such we name it the "cleavage capacity". In **Suppl. Note 5.S2**, we show that the cleavage dynamics follow a nonlinear ODE,

$$\Gamma' = k^1_{\text{PAM}} \left(C_0 + k_{\text{prod}} \cdot t - \sum_j x^0_j (1 - \exp\left(-p_{\text{clv},j} \cdot \Gamma\right)) \right) - k_{\text{degr}} \cdot \Gamma, \qquad \Gamma(0) = 0, \tag{5.7}$$

where C_0 is the initial RNP copy number at $t = 0$. Again, we consider not individual sites i, but classify them according to mismatch pattern j. From a solution for $\Gamma(t)$, the free RNP count $C(t)$ and the intact site count $x_j(t)$ of pattern j can be derived as

$$C(t) = \Gamma'(t)/k^1_{\text{PAM}} \tag{5.8}$$

$$x_j(t) = x^0_j \exp(-p_{\text{clv},j} \cdot \Gamma(t)). \tag{5.9}$$

Equation 5.7 shows that cleavage capacity Γ is generated by initial RNP presence, $k^1_{\text{PAM}} C_0$, and by RNP production, $k^1_{\text{PAM}}\, k_{\text{prod}} \cdot t$, while Γ is limited by RNP degradation, $k_{\text{degr}} \cdot \Gamma$. Finally, the last term in parentheses indicates that Γ is limited as a result of cleaving DNA sites, $k^1_{\text{PAM}} \sum_j \left(x^0_j - x_j(t)\right) = k^1_{\text{PAM}} \sum_j x^0_j \left(1 - \exp(-p_{\text{clv},j} \cdot \Gamma)\right)$. This term reflects that Cas9 complexes are depleted from the pool of active RNPs as they remain stably DNA-bound after cleavage. Although irreversible depletion of RNPs after cleavage is a simplification of reality, where other DNA-binding proteins such as polymerases may displace and recycle it, we assume it here as an extreme limit of DSB absorption. If a genome contains enough sites that allow cleavage (large x^0_j and $p_{\text{clv},j}$ values), this DSB-absorption effect could accelerate RNP depletion beyond protein degradation, and hence determine the dynamics of Γ. One can regard this limit as one where all genomic sites "compete" for available RNP: the first few DSBs deplete RNP levels and hence slow down subsequent DSB dynamics. As a result, the cleavage dynamics on one site become dependent on the other sites in the genomic environment, considerably complicating the matter of predicting DSB dynamics.

We investigate the relevance of this DSB-absorption effect in different contexts. RNP absorption to DSBs is strongest around $t = 0$ and $\Gamma = 0$ as all sites i are intact. Its later magnitude is bounded from above by its linear-order dependence on Γ,

$$k^1_{\text{PAM}} \sum_j x^0_j (1 - \exp\left(-p_{\text{clv},j} \cdot \Gamma\right)) \leq k^1_{\text{PAM}} \sum_j x^0_j\, p_{\text{clv},j} \cdot \Gamma = k^{\text{max}}_{\text{absorp}} \cdot \Gamma. \tag{5.10}$$

This early-time DSB absorption rate $k^{\text{max}}_{\text{absorp}} \cdot \Gamma$ should be compared to the protein degradation term $k_{\text{degr}} \cdot \Gamma$. With CRISPRzip, we predict $p_{\text{clv},j}$ for all 137,980 mismatch patterns j with up to 7 mismatching base pairs at varying negative supercoiling levels σ. These calculations require a value for k_{unb}, which is strongly context-dependent. Single-molecule tracking studies in *E. coli* estimated $k_{\text{unb}} = 60 \pm 13\,\text{s}^{-1}$ and

$k_{\text{bnd}} = 40 \pm 12\,\text{s}^{-1}$ [11], while experiments in mice estimated $k_{\text{unb}} = 1.33 \pm 0.18\,\text{s}^{-1}$ [9]. We consider these two scenarios as examples of target search on prokaryotic and eukaryotic genomes. Next, we calculate site count $x_j^0 = N \cdot p_j$ of pattern j, with the pattern probability p_j as in **Equation 5.4** and the PAM count N on a single haploid genome. The genome copy number does not affect DSB absorption, as it equally increases x_j^0 and decreases k_{PAM}^1 (**Equation 5.1**). Finally, we calculate k_{PAM}^1 using the values for k_{unb} and k_{bnd} as listed above. For the eukaryotic genome, we assume PAM binding is similar to or faster than in prokaryotes, such that $k_{\text{bnd}} \gg k_{\text{unb}}$. In this limit, all RNPs are effectively always PAM-bound and target search time depends only on PAM-unbinding rates (**Equation 5.1**). Our calculation of k_{PAM}^1 considers delay due to trapping based on our previous predictions for μ_j (**Figure 5.1C**).

Our calculations show that DSB absorption effects are generally stronger for smaller genomes (small N) and have a complex dependence on σ (**Figure 5.3B**). Large PAM copy numbers N reduce DSB absorption by slowing down target search, but they also increase the off-target site count, which increases DSB absorption on strongly negatively supercoiled genomes. While strong negative supercoiling σ does promote off-target cleavage, it also causes target search delay due to trapping, resulting in reduced DSB absorption.

Based on these values for $k_{\text{absorp}}^{\text{max}}$, we expect that RNP depletion is unaffected by DSB absorption in human cells. The hg38 human reference genome [45] contains 147 million PAM sites (nuclear DNA, both strands). Assuming that 2–3% of the genome is available for protein binding [46], we estimate a total of $N = 3.7 \cdot 10^6$ Cas9-accessible PAM sites per haploid genome. RNP degradation takes place around $k_{\text{degr}} \approx 10^{-5}\,\text{s}^{-1}$ according to experimental work indicating half-lives of 10-20 hours in human cell lines [47, 48]. Compared to such degradation rates, the estimated DSB absorption rates can safely be ignored at any superhelical density (**Figure 5.3B**), in particular considering that the absorption effect is mitigated at long times and for displaced Cas9. Accordingly, we conclude that Cas9 interacts with its target sites independent of genome sequence context, and that competition dynamics need not be considered to predict cleavage.

5.2.4. Cleavage capacity of Cas9 delivery determines editing outcomes

We have established a simple model of Cas9 activity on the human genome, in which the sequence content of the genetic environment has a negligible influence on the DSB dynamics. Here, we use this model to study Cas9 delivery strategies in human gene editing applications. Two strategies are considered, expression of exogenous Cas9 and delivery of pre-assembled RNPs, knowing that these delivery methods lead to different editing dynamics [47].

We predict average Cas9 activity by solving **Equation 5.7**, and we quantify the variation in activity by generating random instances of genomes with different off-target contents (**Methods 5.4**). RNP delivery is characterized by nonzero C_0 and Cas9 expression by nonzero k_{prod}. We predict the cleavage probabilities $p_{\text{clv},j}$ for all patterns j with up to 7 mismatches with CRISPRzip (**Methods 5.4**). These calculations require a value for σ, however, eukaryotic genome organization is complex and dynamic, and the

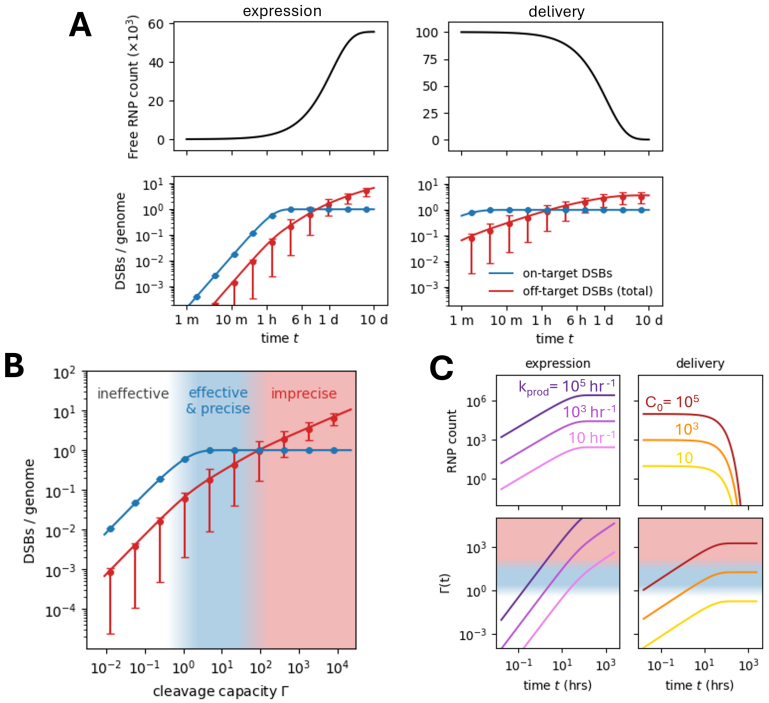


Figure 5.4: Cleavage capacity of Cas9 delivery determines editing outcomes. A Predicted editing outcomes of Cas9 assays with either expressed Cas9 and gRNA ($k_{\text{prod}} = 2 \cdot 10^3\,\text{hr}^{-1}$) or delivered RNP ($C_0 = 10^5$). The two delivery strategies affect the time-evolution of free RNPs (top panels) and hence result in different on-target DSB (blue) and total off-target DSB (red) dynamics. Lines show average predicted DSBs per genome, dots and error bars show mean and standard deviation of 1000 simulations on random genomes. **B** The cleavage capacity Γ unifies the results of the different delivery strategies, and makes general predictions for the precision and effectiveness of an assay in terms of the total on- versus off-target DSBs. **C** Evolution of Γ due to different delivery strategies at variable production rates k_{prod} and delivered RNP copy numbers C_0. Effective and precise on-target cleavage can be attained in an appropriate assay with the right duration.

distribution of superhelical densities is unknown. As Cas9 activity is biased towards (temporarily) underwound regions [22], we assume a hypothetical superhelical density of $\sigma = -0.035$, expecting that such uniformly high superhelicity should overestimate realistic Cas9 activity. The copy numbers x_j^0 and PAM-binding rate k_{PAM}^1 are calculated using $N = 3.7 \cdot 10^6$ PAMs per haploid genome, two genome copies per cell (G_0 or G_1 cell state), $k_{\text{unb}} = 1.33\,\text{s}^{-1}$ [9], and $k_{\text{bnd}} \gg k_{\text{unb}}$.

Our results show that exogenously expressed Cas9 (at a production of $k_{\text{prod}} = 2 \cdot 10^3$ molecules hr^{-1}) accumulates until a steady level is reached, giving rise to increasingly more expected off-target DSBs at long times (**Figure 5.4A**). Delivered pre-assembled RNPs (initial copy number $C_0 = 10^5$) cause the RNP level to be high

initially but steadily decrease over the course of hours to days. Both simulations eventually attain a high degree of off-targeting due to the strongly underwound genome.

The dynamics of cleavage vary with delivery method, but the editing outcomes do not. **Equation 5.9** shows that in a given genome, there is a fixed mapping from Γ to $x_j(t)$. When two Cas9 assays have an identical cleavage capacity Γ, they induce the same predicted DSBs. Hence, instead of predicting cleavage dynamics for particular experiments, we calculate universal editing outcomes as a function of Γ. For our representation of the human genome, assays with a cleavage capacity of $\Gamma \ll 1$ are ineffective on the target site, while assays with $\Gamma \gg 100$ give rise to imprecise cleavage (**Figure 5.4B**). Inside the range of Γ values where edits are effective and precise, one can even optimize for maximum on-target success while keeping off-target activity minimal, corresponding to $\Gamma \approx 10$ in our genome. To achieve such effective and precise edits, an appropriate delivery method should be chosen and tuned (**Figure 5.4C**). In this respect, delivering pre-assembled RNP may be the preferred strategy thanks to its temporary nature, yielding steady editing outcomes without a need for temporal control.

5.2.5. Cleavage is dominated by few strong off-targets

As the cleavage capacity Γ connects the editing outcomes of various Cas9 assays, we use it to investigate which genomic sites cause off-target activity. Studies of off-target activity by Cas9 are typically limited to the identification of "strong" off-targets that are likely to get cleaved, but genomes contain a much greater number of weaker off-targets. It is hard to determine the extent of Cas9 background cleavage on these weak off-targets, as the individual sites are often cleaved too rarely to be detectable. State-of-the-art *in vivo* cleavage assays like GUIDE-Seq [2] rely on next-generation sequencing (NGS) and accordingly have DSB sensitivities limited by the 0.1% error rate of NGS platforms [49]. In addition, the detection of rare DSBs is limited by the finite number of cells or genomes that can be sampled in an experiment. With our model, we predict all target contributions to the full DSB profile on a human genome.

First, we review the cleavage probabilities $p_{\text{clv},j}$ and presence p_j of mismatch patterns j as obtained previously for the human genome under uniform high supercoiling $\sigma = -0.035$ (**Figure 5.5A**). Categorizing the patterns by their mismatch number reveals opposing trends between $p_{\text{clv},j}$ and p_j, illustrating a trade-off between the few strong and many weak off-targets. For instance, when comparing off-targets with 4 mismatches to those with 1 mismatch, the former occur 10^4 times more often while also getting cleaved 10^4 times slower on average.

We inspect the off-target DSBs that were previously predicted at variable Γ (**Figure 5.4B**) and see that the peak in the DSB profile shifts with Γ (**Figure 5.5B**). As Γ increases, relatively weak off-targets are more likely to get cleaved. Because the weaker targets are generally present in greater numbers, they come to dominate the DSB profile when they are cleaved.

The profiles also visualize how much of all cleavage activity is due to rare DSBs, here defined as having a $< 1\%$ cleavage probability at a certain value of Γ (**Figure 5.5B**). Such background activity could remain unobservable, depending on the experimental detection limit. Our model predicts that rare DSBs contribute significantly to the total expected off-target DSBs only in assays with small cleavage capacities ($\Gamma = 0.1$) and

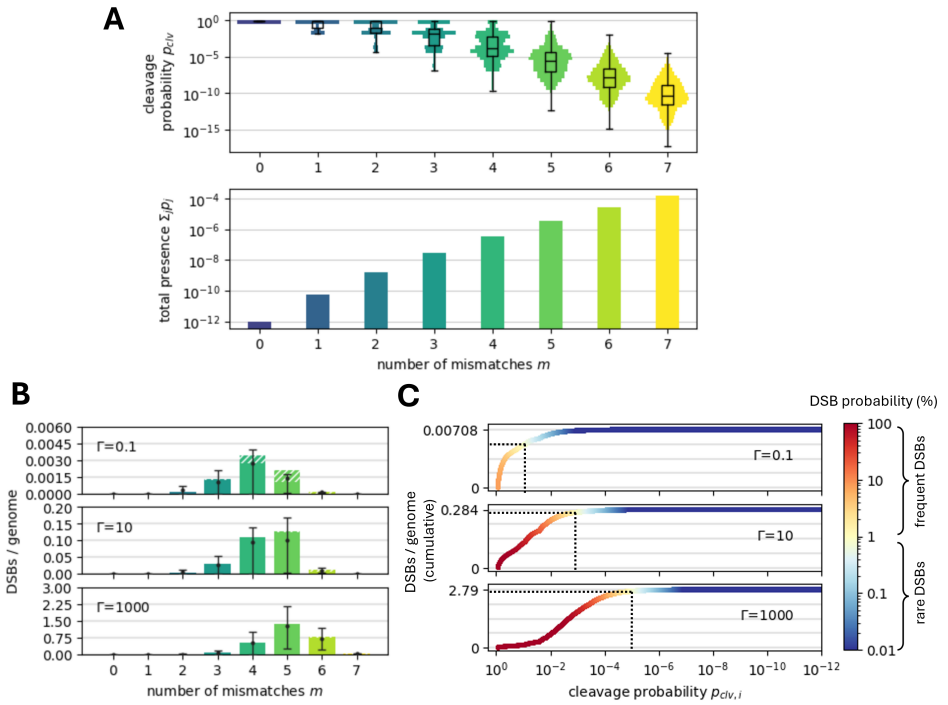


Figure 5.5: Cleavage is dominated by few strong off-targets. A Cleavage probabilities $p_{\text{clv},j}$ and total presence $\sum_j p_j$ of patterns j with 0-7 mismatches. Box plots show median, 25th and 75th percentiles (box) and extremes (whiskers). **B** The distribution of off-target DSBs due to various cleavage capacities Γ. DSBs with $< 1\%$ probability to occur are highlighted in white. **C** Cumulative distribution of DSBs as a function of individual $p_{\text{clv},j}$ values at various cleavage capacities Γ. Colors indicate the probability of a single target i to be cut in an assay with cleavage capacity Γ. Dashed lines separate the contribution due to rare and frequent DSBs.

with low total off-target activity. Rare DSBs rapidly become irrelevant at larger cleavage capacities. The extent of background activity is best illustrated by plotting how DSBs on individual targets i contribute to the total DSB count (**Figure 5.5C**). These plots show the cumulative DSB distribution, highlighting the contributions of frequent DSBs ($> 1\%$ DSB probability, red) and rare DSBs ($< 1\%$ DSB probability, blue). At a small cleavage capacity $\Gamma = 0.1$, rare DSBs collectively make up a fourth of all non-target DSBs, but in this context the overall probability for any off-target DSB to occurs on a genome is small. At larger Γ their contribution vanishes compared to frequent DSBs.

Together, these general predictions of Cas9 cleavage activity show that DSB patterns emerge from a balance between off-target strength and presence. The exact balance follows from the cleavage capacity Γ, so it evolves with time and depends on the particular experimental setup. Even for a genome with homogeneously strong supercoiling, we observe no evidence of significant background cleavage unless off-target activity is low overall. As such, we expect that in general off-target activity

takes place predominantly on strong off-targets that can be identified experimentally.

5.3. Discussion

This chapter presents a kinetic description of Cas9 activity *in vivo* that connects heterogeneous target recognition to genome-scale target search and cleavage dynamics. We predict that the sequence composition of a genome does not considerably affect the expected cleavage dynamics of Cas9 *in vivo*. Although partially matching non-targets can in principle delay target search, and strong off-targets can absorb Cas9 after cleavage, we expect that neither effect is relevant under physiological conditions in human cells. As a consequence, Cas9 operates effectively independently on different genomic sites, allowing cleavage dynamics to be predicted without explicit consideration of genetic background sequence. Also, we predict that one can safely neglect the possibility of large-scale background activity beyond the detection limit of cleavage assays. These simplifications enable a direct translation of *in vitro*-derived R-loop kinetics to *in vivo* settings. For such *in vivo* DSB predictions, one can employ the cleavage capacity as a single abstract quantity to unify Cas9 editing outcomes across delivery strategies and experimental implementations.

Some biological elements could be added to our framework to provide a more complete description of Cas9 activity in cells. First, chromatin state and accessibility could be considered more explicitly. Eukaryotes have particularly complex genome organization [46], which is challenging to model appropriately. In our random-genome model, we assumed that 2–3% of the eukaryotic genome is available for binding. As our model predicts average Cas9 activity, there is no need to consider the accessibility of particular genes across cell types and cell states, as long as the total genome accessibility is approximately constant. However, genome organization is highly dynamic, making different DNA sites available at different times. Also, genome accessibility is partially controlled through the superhelical state of DNA, which can speed up or slow down R-loop formation by Cas9 [43]. Moreover, in regions that are transcribed or replicated, DNA- and RNA-polymerases locally and temporally induce superhelicity and force Cas9 to unbind from its DNA target [10].

A second relevant aspect is DNA repair. This process reverts DSBs, releasing Cas9 to repeat its target search, and sometimes successfully restoring the original sequence of the site. A simple DNA repair model could be an exponential process with characteristic rate k_{repair} and a success rate of p_{repair} [50]. We expect such a model to generate kinetic competition between cleavage and repair, during which erroneously repaired DNA sites gradually accumulate. However, the repair pathway strongly depends on organism, cell state and access to the DSB, and as the dynamics and mutation rate of DSB repair of these pathways vary greatly, we should not expect to find general values for k_{repair} and p_{repair} [51–53].

Third, one could consider the effects of genome replication and cell division. In *E. coli*, these processes may take place on a timescale that is similar to or shorter than target search by Cas9, even more so compared to the timescale on which mutations accumulate due to inaccurate DSB repair. Genome replication and cell division respectively change the copy numbers of DNA sites and Cas9 complexes with a factor 2

and 1/2, so we do not expect them to drastically affect the total cleavage dynamics. Moreover, when considering multiple generations, genomic mutations due to DSB repair are inherited and affect the genomic context in which Cas9 operates.

Along the above suggestions for model expansions, we indicate the need for experimental studies of Cas9 target search in eukaryotes, on which the quantitative power of our model relies. To our knowledge, only one single-molecule tracking study was carried out in eukaryotic cells [9], providing less insight into DNA-binding and -unbinding by Cas9 than offered by the multiple studies in *E. coli* [10–12]. Also, few studies exist that quantitatively report on the distribution of superhelical density in eukaryotic genomes [41, 43]. However, even if more quantitative figures were available, making general predictions about eukaryotes remains challenging as a result of their great variability and complexity.

This work provides a concrete example of how molecular models of CRISPR enzymes can be scaled to predict genome-wide activity *in vivo*. Using CRISPRzip, a mechanistic target recognition model that was trained on *in vitro* data, we connect detailed R-loop kinetics to cellular editing outcomes. This strategy is readily extendable to other CRISPR nucleases and engineered Cas9 variants, including Cas12a, Cascade, Cas9-HF1 and SpRYCas9. As such, the framework presented here highlights a general route toward integrating biophysical insight with practical genome editing applications.

5.4. Methods

5.4.1. Sequence-averaged CRISPRzip predictions

The CRISPRzip protocol to acquire a new parameter set is used to obtain sequence-averaged parameters (**Chapter 3**). We take the original two gRNAs and DNA target libraries (gRNA 1&2 and DNA library 1&2) and perform a 90/10 split to obtain training and test sets. Next, we perform 200 optimization runs to train the model parameters ($U_b^{\mathsf{P,pt}}$, Q_b^{P} and k_{f}) on the libraries simultaneously while excluding sequence effects (by setting the R-loop scaling coefficient $\alpha = 0$). We select the top-50% performing solutions from the optimization results and calculate their median landscape parameters ($U_b^{\mathsf{P,pt}}$, Q_b^{P}). Fixing this median landscape, we optimize k_{f} again using the original training set. In all training runs, the parameters perform as well on the test sets as on the training sets. The resulting model parameters are shown in **Figure 5.S1**.

To account for the effects of superhelicity, we adopt the original CRISPRzip parameter values $\theta = 34^\circ$ and $\varphi = 50^\circ$ (**Chapter 3**).

5.4.2. Off-target cleavage probabilities

With the sequence-agnostic CRISPRzip parameter set (**Figure 5.S1**) , we calculate the cleavage probability $p_{\mathsf{clv},j}$ on mismatch pattern j with the equation [54]

$$p_{\mathsf{clv},j} = \left(1 + \sum_{n=0}^{20} \prod_{b=0}^{n} \frac{k_b^{\mathsf{b}}}{k_b^{\mathsf{f}}}\right)^{-1}. \tag{5.11}$$

Here, the microscopic forward and backward rates k_b^{f} and k_b^{b} respectively quantify the rates of R-loop extension from b to $b+1$ hybrid bps and R-loop recession from b to $b-1$ hybrid bps.

5.4.3. Prediction of genome cleavage

We calculate the expected intact target count $x_j(t)$ of mismatch pattern j by solving the ODE presented in **Equation 5.7**, derived in **Suppl. Note 5.S2**. The values $p_{\mathsf{clv},j}$ are predicted by CRISPRzip at a specific superhelical density σ, and the target copy numbers $x_j^0 = N \cdot p_j$ depend on the PAM copy number N and a pattern's expected presence p_j. Solving the ODE with the values $p_{\mathsf{clv},j}$ and x_j^0 yields average expected cleavage dynamics, considering every possible genome.

Next, we generate random genome instances to quantify variability. We perform binning on the collection of all patterns j on the basis of their $p_{\mathsf{clv},j}$ value. The 100 bins B_k have associated values $\bar{p}_{\mathsf{clv},k}$ that are logarithmically spaced with 5 values per order of magnitude in the range $[10^{-20}, 10^{0}]$ s^{-1} . A random genome is obtained by drawing a random site copy number $\bar{x}_k^0 \sim \mathrm{Poisson}(N \sum_{j \in B_k} p_j)$ from each bin B_k. For that genome instance, we obtain cleavage dynamics by solving the ODE for Γ. This process is repeated for 1000 random genomes.

5.5. Supplementary Notes

5.S1. PAM arrival time with non-target traps

We find the average FPT at any individual PAM site by considering all possible state trajectories in Laplace space. On a site i, the distributions of PAM binding, PAM unbinding, seed formation and seed rejection are given by

$$\phi_{\mathsf{bnd}}(s) = \frac{1}{N} \frac{k_{\mathsf{bnd}}}{k_{\mathsf{bnd}} + s}, \tag{5.S1}$$

$$\phi_{\mathsf{unb},i}(s) = \frac{k_{\mathsf{unb}}}{k_{\mathsf{unb}} + k_{\mathsf{seed},i} + s}, \tag{5.S2}$$

$$\phi_{\mathsf{seed},i}(s) = \frac{k_{\mathsf{seed},i}}{k_{\mathsf{unb}} + k_{\mathsf{seed},i} + s}, \tag{5.S3}$$

$$\phi_{\mathsf{reject},i}(s) = \frac{k_{\mathsf{reject},i}}{k_{\mathsf{reject},i} + s}. \tag{5.S4}$$

We obtain the distribution describing how long Cas9 dwells on a site i, $\psi_{\mathsf{unb},i}(s)$. If Cas9, after binding to the PAM, establishes and rejects a partial R-loop exactly n times before unbinding, the convolution of these $2n+1$ exponential processes gives the FPT distribution at the solution state $\psi_{\mathsf{unb},i}^n$. In Laplace space, we multiply the individual processes,

$$\psi_{\mathsf{unb},i}^n(s) = \left(\phi_{\mathsf{seed},i} \cdot \phi_{\mathsf{reject},i}\right)^n \cdot \phi_{\mathsf{unb},i}\,. \tag{5.S5}$$

We consider any number n of PAM unbinding and rebinding cycles and sum over n,

$$\psi_{\text{unb},i}(s) = \sum_{n=0}^{\infty} \psi_{\text{unb},i}^{n}(s) = \sum_{n=0}^{\infty} \left(\phi_{\text{seed},i} \cdot \phi_{\text{reject},i}\right)^{n} \cdot \phi_{\text{unb,i}} = \frac{\phi_{\text{unb},i}}{1 - \phi_{\text{seed},i} \cdot \phi_{\text{reject},i}}. \quad \text{(5.S6)}$$

Now we calculate the FPT distribution of arriving at a particular PAM site. We consider any number n of excursions to other PAM sites.

$$\Psi(s) = \sum_{n=0}^{\infty} \left(\sum_{i} \phi_{\text{bnd}} \cdot \psi_{\text{unb},i} \right)^{n} \cdot \phi_{\text{bnd}} = \frac{\phi_{\text{bnd}}}{1 - \phi_{\text{bnd}} \cdot \sum_i \psi_{\text{unb},i}}. \quad \text{(5.S7)}$$

The FPT distribution can in theory written as a polynomial fraction $\Psi(s) = P(s)/Q(s)$ where $Q(s)$ is a polynomial of the 4th degree. From this, we know that the FPT in temporal space $F(t) = \mathcal{L}^{-1}[\Psi(s)]$ is a mixed exponential distribution with four modes, each with a rate corresponding to a negative root of $Q(s)$. Although a closed form can theoretically be formulated in terms of the solution to the quartic equation $Q(s) = 0$, it is prohibitively long and complex.

To obtain the average FPT of a single Cas9 complex at an individual PAM, $\langle T^1_{\text{PAM}} \rangle$, we make use of the relation $\langle T \rangle = -d\Psi/ds\big|_{s=0}$ for calculating the first moment $\langle T \rangle$ in Laplace space. For the average FPT, this is

$$\langle T^1_{\text{PAM}} \rangle = \frac{1}{k_{\text{bnd}}} + \sum_{j=0}^{N} \left(\frac{1}{k_{\text{bnd}}} + \frac{1}{k_{\text{unb}}} + \frac{k_{\text{seed},j}}{k_{\text{reject},j}} \frac{1}{k_{\text{unb}}} \right) \quad \text{(5.S8)}$$

$$\approx N \left(\frac{1}{k_{\text{bnd}}} + \frac{1}{k_{\text{unb}}} \left(1 + \frac{1}{N} \sum_{j=0}^{N} \frac{k_{\text{seed},j}}{k_{\text{reject},j}} \right) \right), \quad \text{(5.S9)}$$

where we made use of the fact that $N \gg 1$.

5.S2. ODE system of genome cleavage

The reaction network in the rare-binding limit can be simulated with an ODE system (**Figure 5.3A**),

$$x_j' = -k_{\text{PAM}} \cdot p_{\text{clv},j}\, x_j = -k^1_{\text{PAM}}\, C \cdot p_{\text{clv},j}\, x_j \quad \text{(5.S10)}$$

$$C' = -k^1_{\text{PAM}} \sum_j C \cdot p_{\text{clv},j}\, x_j + k_{\text{prod}} - k_{\text{degr}} C \quad \text{(5.S11)}$$

$$x_j(0) = x_j^0 \quad \text{(5.S12)}$$

$$C(0) = C_0 \quad \text{(5.S13)}$$

Here, every mismatch pattern j has an intact site count of x_j and initial site count of x_j^0 and similar for the expected RNP copy number C. RNP is being produced according to the rate k_{prod} an degraded according to the (unimolecular) rate coefficient k_{degr}. The rate $k^1_{\text{PAM}} = k_{\text{PAM}}/C$ is the (bimolecular) rate coefficient of PAM-binding, corresponding to the average PAM-binding rate per RNP: $1/\langle T^1_{\text{PAM}} \rangle$ as defined in **Equation 5.1**. These

chemical rate equations are valid as long as sufficient RNPs $C \gg 1$ are present.

In the ODE for x_j, we bring all target-dependent terms to one side, integrate from 0 to t and define the site-independent quantity $\Gamma(t)$,

$$\frac{x_j'}{p_{\text{clv},j}\, x_j} = -k^1_{\text{PAM}}\, C \tag{5.S14}$$

$$\frac{\log(x_j/x_j^0)}{p_{\text{clv},j}} = -k^1_{\text{PAM}} \int_0^t C\, d\tau = -\Gamma(t). \tag{5.S15}$$

The quantity $\Gamma(t)$ describes the cleavage capacity of Cas9 in the assay. It follows from the anti-derivative of the Cas9 count $\Gamma(t) = k^1_{\text{PAM}} \cdot \int_0^t dt'\, C(t')$, so by construction, its initial condition is $\Gamma(0) = 0$. From a solution for Γ, one obtains the copy numbers of targets and Cas9 complexes with

$$x_j = x_j^0\, \exp(-p_{\text{clv},j} \cdot \Gamma). \tag{5.S16}$$

The full system can be captured in a single ODE for Γ,

$$\Gamma' = k^1_{\text{PAM}} C = k^1_{\text{PAM}} \left(C_0 + \int_0^t C'\, d\tau \right) \tag{5.S17}$$

$$= k^1_{\text{PAM}} \left(C_0 + \int_0^t \sum_j x_j'\, d\tau + \int_0^t k_{\text{prod}}\, d\tau - \int_0^t k_{\text{degr}} C\, d\tau \right) \tag{5.S18}$$

$$= k^1_{\text{PAM}} \left(C_0 + \sum_j \int_0^t x_j'\, d\tau + \int_0^t k_{\text{prod}}\, d\tau - \int_0^t k_{\text{degr}} \frac{\Gamma'}{k^1_{\text{PAM}}}\, d\tau \right) \tag{5.S19}$$

$$= k^1_{\text{PAM}} \left(C_0 - \sum_j (x_j^0 - x_j) + k_{\text{prod}} \cdot t \right) - k_{\text{degr}}(\Gamma - \Gamma(0)) \tag{5.S20}$$

$$= k^1_{\text{PAM}} \left(C_0 - \sum_j x_j^0 (1 - e^{-p_{\text{clv},j}\Gamma}) + k_{\text{prod}} \cdot t \right) - k_{\text{degr}}\Gamma. \tag{5.S21}$$

5.6. Supplementary Figures

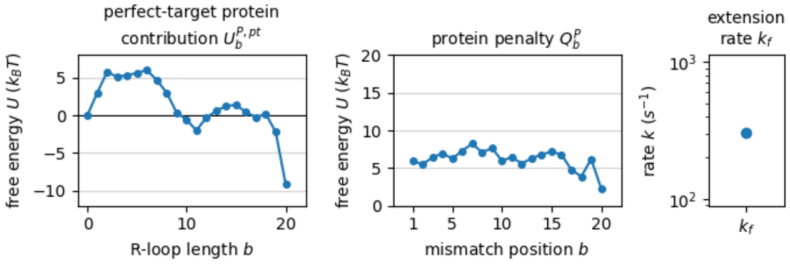


Figure 5.S1: Sequence-averaged CRISPRzip parameters. Sequence-averaged landscape parameters, obtained as explained in **Methods 5.4**.

References

1. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Research* **50,** 3616–3637. doi:10.1093/nar/gkac192 (2022).
2. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology* **33,** 187–197. doi:10.1038/nbt.3117 (2015).
3. Yan, W. X. et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications* **8,** 15058. doi:10.1038/ncomms15058 (2017).
4. Wienert, B. et al. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science* **364,** 286–289. doi:10.1126/science.aav9023 (2019).
5. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology* **17,** 148. doi:10.1186/s13059-016-1012-2 (2016).
6. Concordet, J.-P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research* **46,** W242–W245. doi:10.1093/nar/gky354 (2018).
7. Zhang, G., Luo, Y., Dai, X. & Dai, Z. Benchmarking deep learning methods for predicting CRISPR/Cas9 sgRNA on- and off-target activities. *Briefings in Bioinformatics* **24,** bbad333. doi:10.1093/bib/bbad333 (2023).
8. Yuan, H. et al. An Overview and Comparative Analysis of CRISPR-SpCas9 gRNA Activity Prediction Tools. *The CRISPR Journal* **8,** 89–104. doi:10.1089/crispr.2024.0058 (2025).
9. Knight, S. C. et al. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350,** 823–826. doi:10.1126/science.aac6572 (2015).

10. Jones, D. L. et al. Kinetics of dCas9 target search in Escherichia coli. *Science* **357,** 1420–1424. doi:10.1126/science.aah7084 (2017).

11. Martens, K. J. A. et al. Visualisation of dCas9 target search in vivo using an open-microscopy framework. *Nature Communications* **10.** Number: 1, 3552. doi:10.1038/s41467-019-11514-0 (2019).

12. Olivi, L. et al. Live-cell imaging reveals the trade-off between target search flexibility and efficiency for Cas9 and Cas12a. *Nucleic Acids Research,* gkae283. doi:10.1093/nar/gkae283 (2024).

13. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67. doi:10.1038/nature13011 (2014).

14. Globyte, V., Lee, S. H., Bae, T., Kim, J.-S. & Joo, C. CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. *The EMBO Journal* **38.** doi:10.15252/embj.201899466 (2019).

15. Yang, M. et al. Nonspecific interactions between SpCas9 and dsDNA sites located downstream of the PAM mediate facilitated diffusion to accelerate target search. *Chemical Science* **12,** 12776–12784. doi:10.1039/D1SC02633J (2021).

16. Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nature Communications* **7,** 1–8. doi:10.1038/ncomms12778 (2016).

17. Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Science Advances* **3.** doi:10.1126/sciadv.aao0027 (2017).

18. Yang, M. et al. The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET. *Cell Reports* **22,** 372–382. doi:10.1016/j.celrep.2017.12.048 (2018).

19. Gong, S., Yu, H. H., Johnson, K. A. & Taylor, D. W. DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell Reports* **22,** 359–371. doi:10.1016/J.CELREP.2017.12.041 (2018).

20. Raper, A. T., Stephenson, A. A. & Suo, Z. Functional Insights Revealed by the Kinetic Mechanism of CRISPR/Cas9. *Journal of the American Chemical Society* **140,** 2971–2984. doi:10.1021/jacs.7b13047 (2018).

21. Ivanov, I. E. et al. Cas9 interrogates DNA in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 5853–5860. doi:10.1073/pnas.1913445117 (2020).

22. Newton, M. D. et al. Negative DNA supercoiling induces genome-wide Cas9 off-target activity. *Molecular Cell* **83,** 3533–3545.e5. doi:10.1016/j.molcel.2023.09.008 (2023).

23. Jones, S. K. et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nature Biotechnology* **39.** ISBN: 4158702006465, 84–93. doi:10.1038/s41587-020-0646-5 (2021).

24. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences* **114,** 5461–5466. doi:10.1073/pnas.1700557114 (2017).

25. Boyle, E. A. et al. Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Science Advances* **7.** doi:10.1126/sciadv.abe5496 (2021).

26. Aguirre Rivera, J. et al. Massively parallel analysis of single-molecule dynamics on next-generation sequencing chips. *Science* **385,** 892–898. doi:10.1126/science.adn5371 (2024).

27. Farasat, I. & Salis, H. M. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLOS Computational Biology* **12,** e1004724. doi:10.1371/journal.pcbi.1004724 (2016).

28. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biology* **19,** 1–13. doi:10.1186/s13059-018-1534-x (2018).

29. Zhang, D., Hurst, T., Duan, D. & Chen, S. J. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 8693–8698. doi:10.1073/pnas.1820523116 (2019).

30. Marklund, E. G. et al. Sequence specificity in DNA binding is mainly governed by association. *Science* **375,** 442–445. doi:10.1126/science.abg7427 (2022).

31. Bisaria, N., Jarmoskaite, I. & Herschlag, D. Lessons from Enzyme Kinetics Reveal Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-Based Genome Editing. *Cell Systems* **4,** 21–29. doi:10.1016/j.cels.2016.12.010 (2017).

32. Shvets, A. A. & Kolomeisky, A. B. Mechanism of Genome Interrogation: How CRISPR RNA-Guided Cas9 Proteins Locate Specific Targets on DNA. *Biophysical Journal* **113,** 1416–1424. doi:10.1016/j.bpj.2017.08.013 (2017).

33. Lu, Q., Bhat, D., Stepanenko, D. & Pigolotti, S. Search and Localization Dynamics of the CRISPR-Cas9 System. *Physical Review Letters* **127,** 208102. doi:10.1103/PhysRevLett.127.208102 (2021).

34. Lu, Q. & Pigolotti, S. Target search in the CRISPR/Cas9 system: facilitated diffusion with target cues. *New Journal of Physics* **26,** 043007. doi:10.1088/1367-2630/ad3841 (2024).

35. Zhang, Z., Lamson, A. R., Shelley, M. & Troyanskaya, O. Interpretable neural architecture search and transfer learning for understanding CRISPR–Cas9 off-target enzymatic reactions. *Nature Computational Science* **3,** 1056–1066. doi:10.1038/s43588-023-00569-1 (2023).

36. Moreb, E. A., Hutmacher, M. & Lynch, M. D. CRISPR-Cas “Non-Target” Sites Inhibit On-Target Cutting Rates. *The CRISPR Journal* **3,** 550–561. doi:10.1089/crispr.2020.0065 (2020).

37. Moreb, E. A. & Lynch, M. D. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nature Communications* **12.** Number: 1, 5034. doi:10.1038/s41467-021-25339-3 (2021).

38. Wang, J. C. Degree of superhelicity of covalently closed cyclic DNA's from *Escherichia coli*. *Journal of Molecular Biology* **43,** 263–272. doi:10.1016/0022-2836(69)90266-6 (1969).

39. Sinden, R. R., Carlson, J. O. & Pettijohn, D. E. Torsional tension in the DNA double helix measured with trimethylpsoralen in living E. coli cells: Analogous measurements in insect and human cells. *Cell* **21,** 773–783. doi:10.1016/0092-8674(80)90440-7 (1980).

40. Zheng, G., Kochel, T., Hoepfner, R. W., Timmons, S. E. & Sinden, R. R. Torsionally tuned cruciform and Z-DNA probes for measuring unrestrained supercoiling at specific sites in DNA of living cells. *Journal of Molecular Biology* **221,** 107–122. doi:10.1016/0022-2836(91)80208-C (1991).

41. Naughton, C. et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nature Structural & Molecular Biology* **20,** 387–395. doi:10.1038/nsmb.2509 (2013).

42. Gilbert, N. & Allan, J. Supercoiling in DNA and chromatin. *Current Opinion in Genetics & Development. Genome architecture and expression* **25,** 15–21. doi:10.1016/j.gde.2013.10.013 (2014).

43. Yao, Q., Zhu, L., Shi, Z., Banerjee, S. & Chen, C. Topoisomerase-modulated genome-wide DNA supercoiling domains colocalize with nuclear compartments and regulate human gene expression. *Nature Structural & Molecular Biology* **32,** 48–61. doi:10.1038/s41594-024-01377-5 (2025).

44. Slutsky, M. & Mirny, L. A. Kinetics of protein-DNA interaction: Facilitated target location in sequence-dependent potential. *Biophysical Journal* **87.** ISBN: 00063495/04/12/4, 4021–4035. doi:10.1529/biophysj.104.050765 (2004).

45. Genome Reference Consortium. *Genome Reference Consortium Human Build 38 (GRCh38)* https://www.ncbi.nlm.nih.gov/grc/human. Reference genome sequence downloaded from UCSC hg38 FASTA, January 2026. 2013.

46. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20,** 207–220. doi:10.1038/s41576-018-0089-8 (2019).

47. Karp, H. et al. Packaged delivery of CRISPR–Cas9 ribonucleoproteins accelerates genome editing. *Nucleic Acids Research* **53,** gkaf105. doi:10.1093/nar/gkaf105 (2025).

48. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Research* **24.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press, 1012–1019. doi:10.1101/gr.171322.113 (2014).

49. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nature Methods* **14,** 607–614. doi:10.1038/nmeth.4278 (2017).

50. Brinkman, E. K. et al. Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks. *Molecular Cell* **70,** 801–813.e6. doi:10.1016/j.molcel.2018.04.016 (2018).

51. Song, B., Yang, S., Hwang, G.-H., Yu, J. & Bae, S. Analysis of NHEJ-Based DNA Repair after CRISPR-Mediated DNA Cleavage. *International Journal of Molecular Sciences* **22.** Number: 12, 6397. doi:10.3390/ijms22126397 (2021).

52. Xue, C. & Greene, E. C. DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing. *Trends in Genetics* **37,** 639–656. doi:10.1016/j.tig.2021.02.008 (2021).

53. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology* **20,** 698–714. doi:10.1038/s41580-019-0152-0 (2019).

54. Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depken, M. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Reports* **22,** 1413–1423. doi:10.1016/j.celrep.2018.01.045 (2018).

6

Towards interpretable and adaptable CRISPR activity prediction

6.1. Conclusions

The work in this thesis shows that a molecular understanding of CRISPR gene editing can help explain the outcomes of CRISPR applications in cells. **Chapter 2** established a kinetic framework connecting *in vitro* and *in vivo* conditions; **Chapter 3** presented CRISPRzip, a sequence-aware biophysical model of Cas9 target interrogation, predicting heterogeneous response to RNP concentration and DNA superhelicity; **Chapter 4** investigated the relation between DNA sequence and torque in new high-throughput experiments; and **Chapter 5** demonstrated that in cells, Cas9 cleavage dynamics can be predicted without considering the surrounding genome sequence. Together, these findings illustrate the trajectory from mechanistic enzyme characterization to activity prediction with explicit consideration of the cellular context. While this thesis pursues this route for wild-type SpCas9 and with emphasis on DNA supercoiling, the biophysical methods presented here can be adapted to explore similar pathways for more enzymes and context variables.

6.2. Outlook

Already now, the CRISPRzip framework in its current form can be used to characterize CRISPR enzymes other than wild-type SpCas9. The work in this thesis outlines a pipeline from high-throughput measurement of a nuclease's activity to inference of its sequence-independent protein contribution to R-loop energy landscapes. The barriers and metastable states in these characteristic protein contributions explain enzymes' intrinsic on- and off-target performance beyond specific sequences. As such, this characterization can complement the current practice of benchmarking enzyme activity based on a limited set of gRNAs and DNA targets.

Beyond the effects of enzyme identity and nucleotide sequence, this work highlights DNA superhelicity as an essential variable in CRISPR nuclease target recognition. Relatively few single-molecule studies have investigated the effects of DNA torque on Cas9 [1–4] and other Cas nucleases [1, 5–8]. This work is the first to systematically address the combined influence of nucleotide sequence and supercoiling, revealing a complex interplay between the two variables. Further systematic inquiry should clarify how variable DNA torque couples to DNA and gRNA sequence to determine CRISPR nuclease dynamics. Even if the torque landscape of Cas9 activity were fully characterized, predictions about *in vivo* activity in eukaryotes are limited by the lack of absolute torque values on their genomes. On a molecular level, extensive research has provided a physical description of bare DNA under torque and tension [9–11], and has quantitatively characterized the relation between DNA topology and nucleosomes [12, 13], polymerases [14], SMC architecture proteins [15], topoisomerases [16], and other relevant proteins [17]. However, genome topology in eukaryotes is measured indirectly and expressed in terms of relative supercoiling [18, 19]. It therefore remains unclear how the biophysics of individual protein-DNA interactions integrate into an overall DNA superhelicity level, varying with space and time. Perhaps Cas9 can be employed as a sensor to report directly on torque, contributing to a more physical description of large-scale chromatin organization.

Finally, the challenge remains to translate mechanistic understanding to practical

CRISPR applications. CRISPRzip is a promising candidate tool for gRNA design, providing quantitative on- and off-target prediction, interpretable sequence dependence, mechanistic integration of environmental conditions, and potential adaptation to other RNA-guided nucleases. However, many determinants of *in vivo* activity are lacking from its framework. Some of these determinants may be captured in physical terms and incorporated into the model, for example gRNA folding, PAM variations, indel mutations, local PAM-shuttling effects, and environmental conditions such as ionicity and temperature. Other determinants on the molecular level include interactions with other DNA-binding proteins, among which nucleosomes, polymerases, and DNA repair proteins, which may be studied in a higher-level kinetic model that integrates both these processes and CRISPRzip. Still, many factors remain to *in vivo* CRISPR editing that can hardly be addressed by mechanistic molecular models, from highly specific sequence effects to dynamic genome organization and DNA repair pathways. Machine learning can address this issue of molecular and cellular complexity more effectively than a purely biophysical model. The complementary strengths of these opposite approaches motivate the development of a hybrid model. Here, CRISPRzip's kinetic predictions would feed into a machine learning architecture that modulates them according to the specific context of the genome and application. Recent work applied such a transfer learning approach, starting from a kinetic, mechanistic model of *in vitro* Cas9 activity and embedding it in a neural network for *in vivo* predictions [20]. Other work demonstrated that relatively small Cas9 activity datasets can suffice to adapt general activity predictions to a specific cellular context [21]. In this way, transfer learning built upon CRISPRzip can enable physiologically relevant yet meaningful and flexible gRNA prediction for current and future CRISPR applications.

References

1. Szczelkun, M. D. et al. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proceedings of the National Academy of Sciences* **111,** 9798–9803. doi:10.1073/pnas.1402597111 (2014).

2. Ivanov, I. E. et al. Cas9 interrogates DNA in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 5853–5860. doi:10.1073/pnas.1913445117 (2020).

3. Aldag, P. et al. Probing the stability of the SpCas9–DNA complex after cleavage. *Nucleic Acids Research* **49,** 12411–12421. doi:10.1093/nar/gkab1072 (2021).

4. Newton, M. D. et al. Negative DNA supercoiling induces genome-wide Cas9 off-target activity. *Molecular Cell* **83,** 3533–3545.e5. doi:10.1016/j.molcel.2023.09.008 (2023).

5. Rutkauskas, M. et al. Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Reports* **10,** 1534–1543. doi:10.1016/j.celrep.2015.01.067 (2015).

6. Shi, Y.-J. et al. DNA topology regulates PAM-Cas9 interaction and DNA unwinding to enable near-PAMless cleavage by thermophilic Cas9. *Molecular Cell* **82,** 4160–4175.e6. doi:10.1016/j.molcel.2022.09.032 (2022).
7. Rutkauskas, M. et al. A quantitative model for the dynamics of target recognition and off-target rejection by the CRISPR-Cas Cascade complex. *Nature Communications* **13.** Number: 1, 7460. doi:10.1038/s41467-022-35116-5 (2022).
8. Aris, K. D. P. et al. Dynamic basis of supercoiling-dependent DNA interrogation by Cas12a via R-loop intermediates. *Nature Communications* **16,** 2939. doi:10.1038/s41467-025-57703-y (2025).
9. Lipfert, J., Doniach, S., Das, R. & Herschlag, D. Understanding Nucleic Acid–Ion Interactions. *Annual Review of Biochemistry* **83,** 813–841. doi:10.1146/annurev-biochem-060409-092720 (2014).
10. Marko, J. F. in *Mathematics of DNA Structure, Function and Interactions* 1st ed. *The IMA Volumes in Mathematics and its Applications* 150, 225–249 (Springer, New York, NY, 2020). ISBN: 978-1-4419-0669-4.
11. Gao, X., Hong, Y., Ye, F., Inman, J. T. & Wang, M. D. Torsional Stiffness of Extended and Plectonemic DNA. *Physical Review Letters* **127,** 028101. doi:10.1103/PhysRevLett.127.028101 (2021).
12. Bancaud, A. et al. Structural plasticity of single chromatin fibers revealed by torsional manipulation. *Nature Structural & Molecular Biology* **13,** 444–450. doi:10.1038/nsmb1087 (2006).
13. Kaczmarczyk, A., Meng, H., Ordu, O., Noort, J. v. & Dekker, N. H. Chromatin fibers stabilize nucleosomes under torsional stress. *Nature Communications* **11,** 126. doi:10.1038/s41467-019-13891-y (2020).
14. Gilbert, N. & Marenduzzo, D. Topological epigenetics: The biophysics of DNA supercoiling and its relation to transcription and genome instability. *Current Opinion in Cell Biology* **92,** 102448. doi:10.1016/j.ceb.2024.102448 (2025).
15. Uhlmann, F. SMC complexes: from DNA to chromosomes. *Nature Reviews Molecular Cell Biology* **17,** 399–412. doi:10.1038/nrm.2016.30 (2016).
16. Schoeffler, A. J. & Berger, J. M. DNA topoisomerases: harnessing and constraining energy to govern chromosome topology. *Quarterly Reviews of Biophysics* **41,** 41–101. doi:10.1017/S003358350800468X (2008).
17. Li, S., Vemuri, C. & Chen, C. DNA topology: A central dynamic coordinator in chromatin regulation. *Current Opinion in Structural Biology* **87,** 102868. doi:10.1016/j.sbi.2024.102868 (2024).
18. Naughton, C. et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nature Structural & Molecular Biology* **20,** 387–395. doi:10.1038/nsmb.2509 (2013).
19. Yao, Q., Zhu, L., Shi, Z., Banerjee, S. & Chen, C. Topoisomerase-modulated genome-wide DNA supercoiling domains colocalize with nuclear compartments and regulate human gene expression. *Nature Structural & Molecular Biology* **32,** 48–61. doi:10.1038/s41594-024-01377-5 (2025).

20. Zhang, Z., Lamson, A. R., Shelley, M. & Troyanskaya, O. Interpretable neural architecture search and transfer learning for understanding CRISPR–Cas9 off-target enzymatic reactions. *Nature Computational Science* **3,** 1056–1066. doi:10.1038/s43588-023-00569-1 (2023).

21. Elkayam, S., Tziony, I. & Orenstein, Y. DeepCRISTL: deep transfer learning to predict CRISPR/Cas9 on-target editing efficiency in specific cellular contexts. *Bioinformatics* **40,** btae481. doi:10.1093/bioinformatics/btae481 (2024).

Acknowledgements

I could never have completed my PhD journey without the many people around me that have inspired, taught and supported me. At the times when research was challenging and frustrating, they helped me to overcome my struggles and carry on. As I later think back to my time in Delft, it will be these colleagues and friends that I remember and cherish.

Martin, I feel very lucky to have had such a passionate, committed and friendly person to supervise me. When you offered me this position towards the end of my MEP in 2021, I had no doubt that working with you would be the best PhD opportunity for me. Despite your many chair obligations, you always made time to meet me, making me feel like I was doing you a favor by allowing you to finally talk science (or discuss BN, society, and philosophy). When my research was progressing well, you cheered for my new results and challenged me to go further. When progress stalled and when I doubted myself, you always expressed a deep confidence that things would turn out all right. You listened to me when I felt lost or down, but you never relented to look ahead with positivity, which has been of tremendous help to me. Martin, thank you so much for all of this. There are many pieces of advice you have given me, from stoicism to family planning. In the years ahead, I will particularly remind myself of your advice to look beyond my worries and enjoy the good things while they last.

Many thanks to **Stephen** and **Ieva** for our very pleasant collaboration. Without your data and your experimentalists' perspectives, my theoretical work would have been a lot less meaningful. The two of you and the rest of the **Jones lab** have made me feel so welcome when I came to visit Vilnius. It was a wonderful experience to be among so many fellow CRISPR-minded people. I would like to thank **Ralf**, **Fabian**, and **Julene**, and **Sebastian**, **Anton**, and **Klaus** for all the nice meetings we've had. In a department where no one conducts a similar research project, it is truly invaluable to find others to discuss your scientific interests. **Chirlmin**, thank you for being my copromotor and for encouraging me to seek contact with your group. Thanks to **Carolien** for our work together and to **Ivo**, **Koushik**, and **Carlos** for teaching me so much about smFRET experiments. Thanks to **Raúl**, **Elviss** and the other members of the **DCC** research software support team for working together on my software – I really enjoyed being part of a team, and I learned a lot in the process.

My dear paranymphs **SaFyre** and **Stefan**, I am so grateful for the warm and social vibes that you have brought to BN. I really admire both of you for your open attitude towards others and your talent to get people together. SaFyre, I remember that your arrival at BN immediately made my daily PhD life a lot brighter. You are kind, adventurous, honest, and you truly care for others. I know few people that can listen as well as you do. At the same time, you are great at sharing your own stories, especially when it comes to sex ed. Thank you for making a cis straight white man like me feel so included when discussing all your favorite queer subjects. Stefan, you are an example to me how easy and fun life can be when you don't overthink it. Thanks to your spontaneous ideas and

your “why not” attitude, I got to impress Bebop with my basic jazz piano skills, and we celebrated my thesis submission with free cocktails at a TU Delft socializing event. And thanks to your initiatives to go to Ameland, Bergamo, and more, all of us have grown closer as friends. Thanks for being so optimistic, relaxed and fun.

As a theorist with no possibility of escaping into the lab, I have been so lucky with fun and helpful office mates. **Rachel**, like a big PhD sister you helped me settle in and find a work groove. I remember you telling us to “not work too hard” every time you left the office at the end of the day – a piece of advice that I kept reminding myself of in later years. You pushed for social activities with the theorists, tea breaks, and weekly Friday drinks, which shaped the social atmosphere that still lives on at BN today. Finally, when you were seriously struggling with your research, you reached out to involve others in your project and set up a new experimental research project. You’re an inspiration to me that you should take your own actions to face challenges. **George**, I was happy to sit next to someone so deeply passionate about theoretical biophysics and programming as you, which I later never experienced again. Thanks for your support for developing and maintaining code, you helped me discover how much fun programming can be. Not as much fun though as the Friday night drinks I remember with you, which always ended up later and drunker than I planned. **Felix**, you have been a true example of a young and dedicated scientist to me, and I’m so happy that you have an assistant professor position in Eindhoven now. I was always glad to take your thoughtful advice, and I really enjoyed your warm and kind energy in the office as well as at borrels. **Ireth**, from the moment you showed up for your interview and we chatted about BN, I was so excited that you would join us. I love your sunny personality and your random Spanish (¡Qué calor!) and Dutch (oke doei) exclamations. With you, I could rant and complain about life until you’d tell me to either chill out or to get back to work and do something about it. I was very sad to see you go last year, but I’m glad that you’re pursuing a career that you’re passionate about in the soil business, and I’m also glad that we have kept seeing each other since. Let’s keep doing so! **Ivy**, I know you’re not much of a talker and neither have I been recently, still I really liked having you in the office. You are friendly, funny, and caring, and I believe you will deliver a great PhD thesis soon! I wish you all the best! **Haochun**, you’re such a cool person with such a unique style. I admire how open-minded and daring you are when it comes to making music and art, and I loved hearing your recordings and your live erhu playing in our jam session. **Lisa**, in my last PhD months I was always happy when I wasn’t sitting in the office alone but you showed up too. You always showed genuine interest in how I was doing, and I really appreciated that. You are not only kind but also determined in your research, so I’m sure that you will have a great PhD time at BN.

Having been in a single-member research lab for most of my PhD, I am so grateful for the larger **BN theory** clique, both for the physics and maths discussions we’d have there, and for the friends I made among my “lab mates”. **Pietro**, you are such a chill person to hang out with. Whenever I could join on Friday nights, I was happy that I could always count on you to be there too. I love that when you really get into a conversation, your Italian-ness starts to show more and more. Thanks for your genuine curiosity in my research and my programming activities, and for being as sincerely interested in how I was doing personally. **Yoldaş**, it’s a pity there were only a few months in which

we worked together at BN, I really enjoyed having you around and discussing European politics, Dutch language, and long-form improv theater. Thanks for all those lunch and coffee breaks. **Olivier**, I'm impressed with how much you've achieved in your PhD so far, you will probably write a very nice thesis to defend. And based on your group meeting presentations, I think you'll do great on the defense too. **Timon**, thanks for adopting me at your group meetings, and also for recruiting me for Soft Matter, it was a lot of work but in the end I did enjoy it. **Jos**, I really enjoyed sharing an interest in classical music with you, and I loved that we learned the Poulenc sonata to perform together at your cultural afternoon. And thanks to all the other BN theorists – **Kim**, **Leander**, **Marianne**, **Merlijn**, **Nina**, **Swareena**, **Tamara**, **Ward** – for all the fun breaks and borrels and all the interesting journal clubs and group meetings.

Irène, thank you for being such a wonderful and fun person. Your generous smiles and your cheerful presence have brought a lot of joy to my work life. Somehow the room always brightens up a little when you're there. I will miss our animated political discussions, volleyball afternoons, spring blossom photoshoots and of course our Toon Tellegen and Petit Nicolas reading sessions. We will finish those books one day! I wish you all the best with your position in France and I hope I can come to visit soon. **James**, you have been such a constant factor throughout my PhD that it has been weird to finish it without you as an office neighbor. Fortunately I was too busy writing to come up with random programming questions that I'd ask you and you wouldn't know the answer to. Thanks for entertaining us at our endless tea breaks with your dry humor and your roaring laughter, and for educating me in cricket, cooking and literature. I remember our production of Rachel's PhD movie and my paranymphship at your defense as highlights of my final PhD year. **Beatriz**, I'm so glad I could always convince you (quite easily) to join me for walks. Blowing off steam together outside has been very valuable to me, and I feel like we've grown a lot closer after braving the Campus South wastelands and the psychological PhD wastelands alike. I wish you a happy end to your research project and a life full of seals and hedgehogs. **Christine**, you really always seem to beam positive energy. Of course, mostly so when talking about our sailing adventures or sharks. (I can only imagine your excitement when you find yourself sailing in shark-infested waters.) And I'm very glad that I got you excited too about slow train travel through the Swiss alps. Will you be BN's new train ambassador? **Marieke**, I was always happy to bump into you in the hallway, kitchen or at a voice training workshop. I think back warmly of the heartfelt conversations we've had. Sometimes, you'd introduce me to philosophical perspectives on biophysics and the study of life, other times, we philosophized about PhD life together. **Nynke**, I'm happy to be getting along with the motorcycle-riding, world-cup-sailing, bad-ass person that you are. It was really nice exchanging PhD struggles as well as wedding planning ideas with you. I hope that Marieke and you will smoothly finish your PhD projects and invite me when you celebrate with the last of your traditional cakes. **Fede**, I love your spirited character. Whether it's about organizing social events, hyping spontaneous karaoke, or distracting ourselves from a nauseous mountain ride with word games, you enthusiastically jump on it. And a big thanks to **Ana**, **Anouk**, **Bert**, **Céline**, **Charu**, **Gerard**, **Iain**, **Jeffrey**, **Luci**, **Marcos**, **Marijn**, **Nicky**, **Nikki**, **Sonam**, and other past and present members of the **Biosoft group** for making me feel like a part of your group.

Thanks to **Alejandro**, **Christos**, **Eva**, **Lucia** and **Nils** for recruiting me as a member of the social committee. By joining you, I got to know so much more of BN than my office corner. I think back very fondly of our hilarious dinners, particularly of Nils' cocktail recipe from his student days and Christos' massive moussaka.

Thanks to my BEP and MEP students: **Nathan**, **Femke** and **Marc**. It was a pleasure to see all of you bring your projects to an end and to guide you in the process. Somehow, the challenges you faced always reminded me of my own, so I guess the learning process went both ways. **Mathilde** and **Kate**, it turns out that orchestra friends are also fun to have around the office for their MEP projects! I was always happy when you sat in my office or met you at breaks and drinks. And learning about your projects and progress made me realize once again the similarities between your and my research struggles. We all made it in the end.

Ik zou niet voorbij kunnen gaan aan de steun die ik heb gekregen van vrienden en familie buiten werk. Door met jullie af te spreken, te eten, muziek te maken, te dansen en te zeilen heb ik me vaak met plezier staande weten te houden als ik het lastig had met mijn onderzoek. Er zijn zó veel mensen te bedanken, de **AP-bikkels** en **Boekenbikkels**, het **VU-Orkest**, de **Lucie**, de **Eetclub**, het **NHJO**, mijn **Buitenkunst**-vrienden, de **Hague Hoppers**, het **Ricciotti**, **Pitch Control**, de **Crew**, het **kerstkoor Hark**, en nog veel meer. Onder hen zijn er een paar dierbare vrienden die ik gedurende mijn hele PhD veel heb gezien. **Huib**, het is altijd fijn om af te spreken met jou: je bent warm, grappig, en je bent altijd zo enthousiast om me te zien. Ik vind het enorm inspirerend hoe openhartig jij kan vertellen over je worstelingen. Het is mooi om te zien hoe het je steeds beter lukt om je eigen vrolijke en unieke draai te geven aan het leven. Nu mijn PhD voorbij is, ga ik ook op zoek naar hoe ik mijn leven in wil richten. Ik weet zeker dat ik daarbij veel van je kan blijven leren. **Hina**, dankjewel voor alle spelletjes, boekenclubs en bankhangsessies. Als we afspreken, straalt je opgewekte en hartelijke persoonlijkheid op me af en word ik meteen vrolijk. Ik ga mijn best doen om regelmatig naar Amersfoort te komen! **Guido**, ik weet dat het niet klopt, maar in mijn herinnering kwamen wij elkaar voortdurend tegen op het perron, en bleef je zo ongeveer elke maand bij ons thuis op de bank slapen. In onze gesprekken neem je me altijd zo makkelijk mee in je enthousiasme, vooral als het over muziek en zeilen gaat. Ik ben dankbaar dat we elkaar ook nu nog regelmatig zien en samen blijven zeilen, zingen en musiceren.

Ook is er een aantal mede-PhD'ers die ik in het bijzonder wil bedanken voor de waardevolle gesprekken over alles wat er mooi en lastig is aan promotieonderzoek doen. **Inge**, je weet dat ik je enorm bewonder om je dappere keuze om te stoppen, en ik ben heel blij met hoe het heeft uitgepakt voor je. Dankjewel voor al onze openhartige treingesprekken en voor het schetsen van perspectief. **Evelijn**, volgens mij hebben wij een totaal andere PhD-ervaring gehad, en toch was het altijd heel fijn om mijn worstelingen met je te delen en de jouwe aan te horen. Ik wens je een fantastische tijd in Barcelona, en ik kijk ernaar uit om je daar op te zoeken en te horen over je nieuwe onderzoeks-ups en -downs. **Amber**, ik ben heel erg onder de indruk van je toewijding aan zowel je onderzoek als aan de opleiding die je coördineert. Onze onderzoekswerelden zijn heel anders, ik vond het heel waardevol en leuk om perspectieven met je uit te wisselen. **Daan**, jij kan zoveel enthousiasme uitstralen als het gaat om natuurkunde, of het nu om je eigen PhD-onderzoek gaat of het mijne. Of om de klinische fysica waar je

nu mee bezig bent, ik ben enorm blij voor je dat je die opleidingsplek hebt bemachtigd. Ik wens jullie alle rust en stabiliteit toe en een prachtige tijd met Emily. **Max**, ik vond het altijd superleuk om updates te horen over jou en je Duitse zebravisjes. Ik denk met veel plezier terug aan mijn bezoek aan Konstanz, maar wat betreft PhD-gesprekken herinner ik vooral ons gesprek in Muiden, waarin je me op het hart drukte dat er ook veel plezier te halen is uit het afronden van een PhD. Je had gelijk! **Josefien**, leuk om ervaringen uit te wisselen over onze heel verschillende promotietrajecten, met name over onze proefschriften en onze verdedigingen die toevallig rond dezelfde tijd vallen. Ik wens je heel veel succes en plezier bij de jouwe, ik kijk ernaar uit.

Ik richt een bijzonder dankwoord aan **Zoë** voor onze fijne gesprekken. Je hebt me geholpen om milder te oordelen over mezelf. Wanneer ik terugkijk op mijn tijd in Delft kan ik dankzij jou een plekje geven aan de tegenslagen en teleurstellingen, en meer voldoening halen uit wat ik wel voor elkaar heb gekregen. Ik ben blij dat je erbij was in de laatste maanden toen ik het project met een goed gevoel kon afsluiten.

Pap en **mam**, dankjewel voor alle steun. Jullie geven er veel om dat ik het fijn heb op mijn werkplek en denken met alle liefde mee als ik in een lastige situatie zit. Ik kan bij jullie terecht als ik het even niet meer weet, en jullie laten me voelen dat jullie achter me staan ongeacht welke keuzes ik maak. In mijn zoektocht naar nieuw werk zullen er nog veel twijfels voorbijkomen, en dan ben ik ongetwijfeld weer erg dankbaar voor jullie luisterend oor.

Lieve **Eline**, natuurlijk is het laatste dankwoord aan jou. Jij hebt me zoveel houvast, troost, afleiding, en perspectief gegeven wanneer het lastig was. En wanneer het goed ging, was je altijd zó blij om te luisteren naar wat ik te vertellen had. Er is niemand die me kan aanhoren, begrijpen en steunen zoals jij dat kan, en daarmee ben je onmisbaar geweest voor het succesvol afronden van mijn PhD. Nu is het tijd voor mij om te zoeken naar nieuw werk. Ik ga daarin ongetwijfeld mezelf weer tegenkomen en mijn zelfvertrouwen soms verliezen. Dan zal ik terugdenken aan die keer dat jij je volle vertrouwen uitsprak dat er ook voor mij een plek bestaat waar ik kan opbloeien. Met jou aan mijn zijde zal ik, vallend en opstaand, die plek vinden. Ik ben heel benieuwd waar deze stappen mij, jou en ons gaan brengen. Ik hou van je, en ik kijk enorm uit naar de komende jaren.

Curriculum Vitæ

Hidde Sieger Offerhaus

11 June 1997	Born in Haarlem, The Netherlands.
Sep. 2009 – Jun. 2015	Secondary Education, Petrus Canisius College, Alkmaar, The Netherlands
Sep. 2015 – Jun. 2019	**BSc in Liberal Arts and Sciences**, Amsterdam University College, Amsterdam, The Netherlands *Thesis:* Fiber Optic Microsphere Lens for Common-Path Optical Coherence Tomography *Supervisor:* Prof. Dr. D. Iannuzzi
Sep. 2019 – Jun. 2021	**MSc in Applied Physics**, Delft University of Technology, Delft, The Netherlands *Thesis:* Fast and Flawless Target Search: Analytically Exploring How CRISPR-Cas9 Architecture Unites Competing Demands *Supervisor:* Dr. S. M. Depken
Sep. 2021 – Mar. 2026	**PhD in Theoretical Biophysics**, Department of Bionanoscience, Delft University of Technology, Delft, The Netherlands *Thesis:* From R-loops to Genomes: Connecting Molecular CRISPR-Cas9 Kinetics to Genome Editing *Promotor:* Dr. S. M. Depken *Coprotomor:* Prof. Dr. C. Joo

List of Publications

Journal articles and preprints

1. Offerhaus, H. S., Jaskovikaitė, I., Jones, S. K. & Depken, M. A mechanistic free-energy model explains and predicts sequence- and context-dependent CRISPR-Cas9 activity. *preprint on bioRxiv.* doi:10.64898/2025.12.12.691775 (2025)

2. Jaskovikaitė, I., Offerhaus, H. S., Vinogradovas, M., Barkauskaitė, U., Depken, M. & Jones, S. K. Supercoiling twists Cas9 off-target discrimination when nicking and cleaving. *preprint on bioRxiv.* doi:10.64898/2025.12.12.691775 (2026)

3. Olivi, L., Bagchus, C., Pool, V., Bekkering, E., Speckner, K., Offerhaus, H. S., Wu, W. Y., Depken, M., Martens, K. J. A., Staals, R. H. J. & Hohlbein, J. Live-cell imaging reveals the trade-off between target search flexibility and efficiency for Cas9 and Cas12a. *Nucleic Acids Research,* gkae283. doi:10.1093/nar/gkae283 (2024)

Software

1. Offerhaus, H. *CRISPRzip* version 1.2.2. Dec. 2025. doi:10.5281/zenodo.17818249

2. Offerhaus, H. *CRISPRzip-tool* version v1.0.0. Dec. 2025. doi:10.5281/zenodo.17818523