UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Unraveling the Mechanism of Activation and Functioning of CRISPR-Cas12a and CRISPR-Cas9-Conjugated Complexes

Permalink

https://escholarship.org/uc/item/2zp5v2p9

Author

Saha, Aakash

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Unraveling the Mechanism of Activation and Functioning of CRISPR-Cas12a and CRISPR-Cas9-Conjugated Complexes

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Bioengineering

by

Aakash Saha

September 2024

Dissertation Committee:
- Dr. Giulia Palermo, Chairperson
- Dr. Valentine Vullev
- Dr. Joshua Morgan

Copyright by
Aakash Saha
2024

The Dissertation of Aakash Saha is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENT

I wish to express my deepest gratitude to Dr. Giulia Palermo, whose exceptional kindness and willingness to mentor me, despite my non-traditional background in the field, have been invaluable. She entrusted me with an extraordinary project, one that quickly became a passion of mine, and generously allowed me the space to learn and grow throughout the process. Her unwavering support, particularly in guiding me through the intricacies of writing scientific papers and grants, has been instrumental in my progress during this doctoral journey.

I extend my heartfelt thanks to all the members of our lab. Dr. Pablo Ricardo Arantes provided essential initial training, patiently introducing me to the diverse methodologies within the field. Dr. Mohd. Ahsan has been a significant guiding presence, frequently collaborating with me on numerous projects. His expertise introduced me to various techniques and was pivotal in driving our work toward successful completion. I am also deeply appreciative of Dr. Souvik Sinha, whose insights have kept me abreast of the latest advancements in the field. Additionally, I would like to acknowledge Dr. Chinmai Pindi and Amun Patel for their camaraderie, making my time in the lab both productive and enjoyable.

I am profoundly grateful to the funding agencies, including the National Institutes of Health (NIH) and the National Science Foundation (NSF), as well as the computational resources provided by Anton-2 from the D. E. Shaw Research group. Their generous support made it possible to undertake and complete such an ambitious project.

I also extend my sincere thanks to my committee members, Dr. Joshua Morgan and Dr. Valentine Vullev, for their steadfast support and regular assessments of my progress. Their constructive feedback has been crucial in helping me stay focused and on course throughout my doctoral studies.

Finally, I owe a deep debt of gratitude to my family and friends, who have been my pillars of strength during this journey. Despite the distance between us, my parents in India have patiently endured my emotional ups and downs, always responding with unconditional love and encouragement, which have been my greatest sources of motivation. My friends—Akash Chatterjee, Sinjini Chatterjee, Kuntal Pal, and Jon Christopher Rodrigues—have been an inseparable part of my life, offering unfaltering support through every phase, ensuring that I never lost heart. Their presence has kept my spirits high, allowing me to forge ahead with confidence.

DEDICATION

This Dissertation is dedicated to my mother, Shampa Saha, for being my guiding beacon, and my father, Nitya Gopal Saha, for his unwavering support.

“I’ve never been a natural,

all I do is try, try, try.”

–Taylor Swift

ABSTRACT OF THE DISSERTATION

Unraveling the Mechanism of Activation and Functioning of CRISPR-Cas12a and CRISPR-Cas9-Conjugated Complexes

by

Aakash Saha

Doctor of Philosophy, Graduate Program in Bioengineering
University of California, Riverside, September 2024
Dr. Giulia Palermo, Chairperson

CRISPR-Cas12a has revolutionized genome editing, offering precise control over genetic modifications while also serving as extraordinarily rapid and reliable diagnostic tools. This thesis provides comprehensive insights into Cas12a through advanced molecular dynamics simulations and experimental approaches. The second chapter involves multi-microsecond molecular dynamics simulations to reveal the allosteric switches governing conformational activation of REC2 and Nuc domains in Cas12a. Taking the investigation forward, the third chapter addresses how Cas12a generates double-strand DNA breaks using its single RuvC nuclease domain. Here, long molecular dynamics and free-energy simulations uncovers the pivotal role of an α-helical lid within the catalytically active RuvC domain. This lid anchors the crRNA:target strand duplex and guides the target strand toward the RuvC core, a mechanism corroborated by DNA cleavage experiments. The fourth chapter further investigates the role of the α-helical

lid by examining R-loop formation using cryo-electron microscopy and advanced free-energy simulations. Structural and dynamic insights reveal that the lid transitions from an unstructured loop at the 5-bp seed state to an α-helical structure at the 16-bp and 20-bp states, accompanied by distinct REC domain rearrangements. This explains the accommodation of the non-target strand (NTS) followed by the target strand (TS) at the active site of Cas12a. The fifth chapter focuses on the *trans*-cleavage property of Cas12a, which is the basis for nucleic acid detection. Kinetic studies, bolstered by thermodynamic analyses, show that the *trans*-cleavage activity rate of Cas12a is significantly enhanced for hairpin DNA structures. This enables faster detection of clinically relevant double-stranded DNA targets with improved sensitivity and specificity. Finally, the sixth chapter investigates CRISPR-Cas9-based Adenine Base Editors (ABEs), which hold significant promise for addressing human genetic diseases caused by point mutations. We identify critical residues and demonstrate that the dimerization of TadA8e (the deaminase domain) and its unique juxtaposition to Cas9 are pivotal for efficient DNA deamination by ABE8e, the most efficient ABE to date. Overall, this thesis advances our understanding of CRISPR-based (Cas12a and Cas9) genome-editing tools, providing mechanistic insights into critical processes that will enrich fundamental knowledge and facilitate further engineering strategies for genome editing and diagnostic applications.

Table of Content

Abstract . . . . . **viii**

Table of Content . . . . . **x**

List of Figures . . . . . **xix**

List of Tables . . . . . **xxvi**

I Introduction **1**

1.1 Overview of CRISPR-Cas systems 1

1.2 CRISPR-Cas12a: A extraordinary genome-editing and nucleic-acid detection tool. 4

1.2.1 Structural Details 4

1.2.2 Guide RNA (crRNA) recognition and processing 5

1.2.3 PAM Recognition and Target DNA Binding 6

1.2.4 Allosteric Activation of Cas12a 9

1.2.5 *cis*- and *trans*-cleavages by Cas12a 10

1.2.6 Off-target *cis*-cleavages in Cas12a 12

1.3 CRISPR-Cas9: The golden genome-editing tool 13

1.3.1 A brief overview 13

1.3.2 A comparative study of CRISPR-Cas9 with respect to CRISPR-Cas12a 14

1.4 Therapeutic Applications of CRISPR-Cas Systems 16

1.5 Diagnostic Applications of CRISPR-Cas12a 17

1.6 Overarching Questions 19

II Molecular Dynamics Reveal a DNA-Induced Dynamic Switch Triggering Activation of CRISPR-Cas12a 21

2.1 Abstract 21

2.2 Introduction 22

2.3 Results 26

2.3.1 Conformational Flexibility of Cas12a Bound to RNA and DNA 26

2.3.2 Large–amplitude Motions and Conformational Ensemble 29

2.3.3 Coupled Motions of Protein Domains 34

2.4 Discussion 38

2.4.1 Dynamic “Switch” upon DNA Binding 38

2.4.2 Conformational Activation for TS Processing 40

2.5 Conclusions 43

2.6 Materials & Methods 45

2.6.1 Structural Models 45

2.6.2 Molecular Dynamics (MD) Simulations 46

2.6.3 Principal Component Analysis (PCA) 47

2.6.4 Cross-Correlation analysis 49

2.6.5 Generalized Correlations analysis based on Mutual Information 50

2.7 Acknowledgments 52

III An Alpha-helical Lid Guides the Target DNA toward Catalysis in CRISPR-Cas12a 53

3.1 Abstract 53

3.2 Introduction 54

3.3 Results 57

3.3.1 An α-helical Lid Anchors the crRNA:TS 57

3.3.2 REC2 and Nuc Facilitate the Traversal of the TS 63

3.3.3 Traversal of the DNA TS and Accommodation in the RuvC Catalytic Pocket 65

3.3.4 Mutations of the Lid Adversely Affect Target Strand Traversal . 70

3.4 Discussion 73

3.5 Materials & Methods 79

3.5.1 Structural Models 79

3.5.2 Molecular Dynamics (MD) Simulations 80

3.5.3 Umbrella Sampling Simulations 82

3.5.4 Alchemical Free Energy Perturbation 83

3.5.5 FnCas12a Expression and Purification 84

3.5.6 FnCas12a Nuclease Activity Assays 86

3.6 Acknowledgements 87

IV Cas12a Domain Flexibility Guides R-loop Formation and Forces RuvC Resetting 88

4.1 Abstract 88
4.2 Introduction 89
4.3 Results 91
4.3.1 Structures of Cas12a on Pathway to Cleavage 91
4.3.2 Cas12a Uses a 5-bp Seed 98
4.3.3 Fast Dissociation Kinetics of Early and Middle R-loop Intermediates Promote Efficient Target Search 101
4.3.4 REC2 Flexibility Accommodates R-loop Extension but Delays Contacts 102
4.3.5 Late REC2 Docking Promotes Catalysis 107
4.3.6 NTS Poised for Cleavage in the Active Site 109
4.3.7 MD Simulations Probe RuvC Exposure during R-loop Formation 113
4.3.8 Dramatic Conformational Changes Precede TS Cleavage and Reset the Active Site 115
4.4 Discussion 117
4.5 Acknowledgements 122
4.6 Materials & Methods 122
4.6.1 Cas12a Cloning and Purification 122
4.6.2 Target DNA Substrates 124
4.6.3 Cryo-EM Sample Preparation, Data Collection 124
4.6.4 Cryo-EM Data Processing 125
4.6.5 Model Building and Refinement 130
4.6.6 Cleavage Time Courses 130

4.6.7 Plasmid Interference Assay 132
4.6.8 Molecular Dynamics Simulations 132
4.6.9 Multiple Walker Adaptively Biased MD 134
4.6.10 Quantification and Statistical Analysis 135

V Enhancement of CRISPR/Cas12a *Trans*-cleavage Activity using Hairpin DNA Reporters 136

5.1 Abstract 136
5.2 Introduction 137
5.3 Materials and Methods 140
5.3.1 Oligonucleotides 141
5.3.2 Heterologous Expression and Purification of FnCas12a 148
5.3.3 DNA Preparation of Plasmid Containing the SARS-CoV-2 Sequence (S gene) 148
5.3.4 Production of SARS-CoV-2 Amplicons through PCR 149
5.3.5 Preparation and Standardization of *Salmonella* Strain 150
5.3.6 PCR Amplification of *Salmonella* Samples 150
5.3.7 Melting Assays and Thermodynamic Analysis 151
5.3.8 Fluorescence *Trans*-cleavage Assays 152
5.3.9 Michaelis-Menten Enzyme Kinetic Measurements and Data Analysis 154
5.3.10 Native Polyacrylamide Gel Electrophoresis (PAGE) Assays . . . 156
5.3.11 PAGE Analysis of *Trans*-cleavage Digestion of FRET-based DNA Reporters 157

5.3.12 Structural Models 157
5.3.13 Molecular Simulations 158
5.3.14 Docking and Binding Energies 159
5.4 Results and Discussion 160
5.4.1 Design of FRET-based Hairpin DNA Reporters 160
5.4.2 Faster and enhanced Signal Change Using Hairpin DNA Reporters 164
5.4.3 Hairpin Reporters Show Enhanced Rate of *Trans*-Cleavage . . . 165
5.4.4 Mechanistic Insight into the Hairpin DNA reporter interactions with Cas12a 169
5.4.5 Enhanced Sensitivity and Specificity using Hairpin DNA Reporter 172
5.5 Acknowledgements 178

VI Dimerization of the Deaminase Domain and its Interactions with Cas9 Boost Base Editing Efficiency in ABE8e 179

6.1 Abstract 179
6.2 Introduction 180
6.3 Results 184
6.3.1 Multiple rounds of directed evolution have stabilized the dimeric state of TadA 184
6.3.2 DNA–TadA–Cas9 interactions boost DNA deoxyadenosine deamination by ABE8e 188
6.3.3 Key substitutions introduced during ABE8e's evolution tune TadA stability for enhancing DNA deamination capability . . . 198
6.4 Discussion 205

6.5 Materials & Methods . . . 209
6.5.1 Structural Models . . . 209
6.5.2 Molecular Dynamics (MD) Simulations . . . 209
6.5.3 Umbrella Sampling Simulations . . . 211
6.5.4 Alchemical Free Energy Perturbation . . . 212
6.5.5 Preparation of ABE-sgRNA Samples . . . 214
6.5.6 Single-turnover *in vitro* DNA and RNA Deamination Assays . . . 214
6.5.7 Thermal Stability Assay using CD Spectroscopy . . . 215
6.5.8 Fluorescence Resonance Energy Transfer (FRET) Measurements 215
6.6 Acknowledgements . . . 216

Appendices . . . 218

A Supporting Information for Chapter II . . . 218

A.1 Structural models . . . 218
A.2 Molecular Dynamics (MD) . . . 219
A.3 Principal Component Analysis (PCA) . . . 220
A.4 Volumetric Analysis . . . 223
A.5 Cross-Correlation Analysis . . . 224
A.6 Generalized Correlation (GC_{ij}) Analysis based on Mutual Information . 225
A.7 Per-domain Correlation Scores (C_s) . . . 226
A.8 Supplementary Figures . . . 228

B Supporting Information for Chapter III . . . 241

B.1 Structural Models . . . 241

B.2 Molecular Dynamics (MD) Simulations 245
B.3 Principal Component Analysis 247
B.4 Umbrella Sampling Simulations 249
B.5 Calculation of Bending Angles 252
B.6 Alchemical Free Energy Simulations 253
B.7 FnCas12a Expression and Purification 255
B.8 FnCas12a Nuclease Activity Assays 256
B.9 Supplementary Tables 257
B.10 Supplementary Figures 274
B.11 Statistics and Reproducibility 292
B.12 Ethics and Inclusion Statement 293
B.13 Data Availability Statement 293

C Supporting Information for Chapter IV 294

C.1 Supplementary Tables 294
C.2 Supplementary Figures 302
C.3 Data and code availability 308
C.4 Bacterial Strains and Growth Conditions 308
C.5 Limitations of the Study 309

D Supporting Information for Chapter V 310

D.1 Supplementary Tables 310
D.2 Supplementary Figures 313

E Supporting Information for Chapter VI 326

E.1 Structural Models 326
E.2 Molecular Dynamics (MD) Simulations 327
E.3 Umbrella Sampling (US) Simulations 328
E.4 Differential Inter-residue Interactions 330
E.5 Alchemical Free Energy Perturbation 331
E.6 Cloning and Protein Expression 333
E.7 Protein Purification 334
E.8 *In vitro* Transcription of Guide RNA 335
E.9 Assembly of ABE-sgRNA Complexes 335
E.10 Single-turnover *in vitro* DNA and RNA Deamination Assays 336
E.11 Circular Dichroism (CD) Measurements 337
E.12 Thermal Stability Assay using CD Spectroscopy 337
E.13 Preparation of Fluorescently-labelled ABEs 338
E.14 Supplementary Tables 339
E.15 Supplementary Figures 342
E.16 Statistics and Reproducibility 359

List of Figures

2.1 Overview of CRISPR-Cas12a. 24

2.2 Cas12a flexibility along the dynamics. 29

2.3 Large-amplitude motions and conformational space adopted by CRISPR-Cas12a. 31

2.4 Correlated motions of CRISPR-Cas12a. 37

2.5 Changes in the RuvC binding groove of CRISPR-Cas12a before and after NTS cleavage. 42

3.1 Structures of CRISPR-Cas12a after DNA non-target strand cleavage. . . 56

3.2 The α-helical lid anchors the crRNA:TS duplex. 61

3.3 Bending of major domains in CRISPR-Cas12a. 64

3.4 Traversal of the DNA target strand toward the RuvC catalytic core and role of the α-helical lid. 70

3.5 Model for the DNA target strand traversal toward the RuvC catalytic core in CRISPR-Cas12a. 78

4.1 Cas12a R-loop intermediates captured by cryo-EM. 90

4.2 Cas12a rapidly checks for DNA complementarity with a 5-bp seed. . . . 99

4.3 Middle R-loop propagation is marked by REC2 flexibility, delaying R-loop contacts and RuvC activation. 105
4.4 Structures of RuvC-mediated DNA cleavage by Cas12a. 111
4.5 Structural model for Cas12a R-loop formation and DNA cleavage. . . . 118

5.1 Schematic for Cas12a *trans*-cleavage activity based on hairpin DNA reporters. 161
5.2 Design and characterization of different hairpin DNA reporters as FRET-based reporters of LbCas12a *trans*-cleavage activity. 163
5.3 Kinetic analysis of LbCas12a *trans*-cleavage activity. 167
5.4 Mechanistic insights from molecular simulations. 170
5.5 Sensitivity and specificity assays. 175

6.1 Overview of Adenine Base Editor's (ABE) structure. 183
6.2 Dimerization of TadA enzymes. 187
6.3 Gain/loss of interaction strength in ABE8e compared with ABE0.1 and ABE7.10. 191
6.4 Key interactions between the docking domain of TadA8e with the DNA and Cas9 for efficient DNA deoxyadenosine deamination by ABE8e. . . 196
6.5 Functional role of amino acid substitutions introduced during directed evolution of ABE8e. 201

A.1 Root Mean Square Fluctuations (RMSF) profiles of the protein. 228
A.2 Root Mean Square Deviation (RMSD) profiles of the protein. 229
A.3 Root Mean Square Deviation (RMSD) profiles of the RNA. 230

A.4 Root Mean Square Deviation (RMSD) profiles of the DNA. 231

A.5 Cumulative contribution of all the principal components. 232

A.6 Volumetric Analysis of the groove between the REC and NUC lobes. . . 233

A.7 Principal Component Analysis. 234

A.8 Cross-Correlation of RNA-bound FnCas12a. 235

A.9 Cross-Correlation of DNA-bound FnCas12a. 236

A.10 Cross-Correlation of DNA-bound FnCas12a'. 237

A.11 Cross-Correlation of DNA-bound AsCas12a. 238

A.12 Cross-Correlation of RNA-bound LbCas12a. 239

A.13 Generalized Correlation Analysis. 240

B.1 Molecular dynamics (MD) simulations of the 6GTG[1] cryo-EM structure of CRISPR-Cas12a (Figure 3.1a). 274

B.2 Structure and dynamics of the complete CRISPR-Cas12a system (Figure 3.1c). 275

B.3 Stability (a) and flexibility (b) of the lid (residues L1008–K1021) along MD simulations. 276

B.4 Molecular dynamics (MD) simulations of the 5NFV[2] X-ray structure of CRISPR-Cas12a (Figure 3.1b). 277

B.5 Bending of major domains and effects on the scissile phosphate along MD simulations of the 5NFV[2] X-ray structure of CRISPR-Cas12a (Figure 3.1b). 278

B.6 Bending of major domains along MD simulations of the complete CRISPR-Cas12a system (Figure 3.1c). 279

B.7 Number of contacts between major domains of Cas12a and nucleic acids. 280
B.8 Overview of the catalytically competent CRISPR-Cas12a built from related X-ray structures. 281
B.9 Molecular dynamics simulations of the catalytically competent CRISPR-Cas12a for TS cleavage. 282
B.10 Free energy profiles for the traversal of the DNA target strand (TS) toward the RuvC catalytic core. 283
B.11 Two-dimensional (2-D) free energy maps for the traversal of the DNA target strand (TS) toward the RuvC core. 284
B.12 Convergence of the free energy profiles obtained through Umbrella Sampling (US) simulations. 285
B.13 Interactions between the lid and the crRNA:TS duplex along the minimum energy path (MEP) from Umbrella Sampling simulations of the wild-type CRISPR-Cas12a. 286
B.14 Lid dynamics along the minimum energy path (MEP) of the Umbrella sampling simulation. 287
B.15 Dynamics of major domains at different states along the free energy profiles of the Umbrella sampling simulations. 288
B.16 Convergence of the relative free energy changes obtained through alchemical free energy simulations (AFE). 289
B.17 *In vitro* DNA TS cleavage assay of Cas12a mutants. 290
B.18 *In vitro* dsDNA cleavage assay of Cas12a mutants. 291
B.19 *In vitro* DNA TS cleavage assay of Cas12a mutants. 292

C.1 Local resolution of intermediate R-loop maps. 302

C.2 Data processing pipelines for Cas12a R-loop datasets. 303

C.3 Cas12a early and middle R-loops. 304

C.4 Cas12a RuvC activation. 306

C.5 Data processing pipeline for Cas12a post-NTS cleavage dataset. 307

D.1 Thermodynamic characterization of the FRET-based hairpin DNA reporters. 313

D.2 Background fluorescence of the DNA reporters. 314

D.3 *Trans*-cleavage activity was tested by using different crRNA and dsDNA activators. 315

D.4 *Trans*-cleavage activity was tested using different orthologs of Cas12a enzyme. 316

D.5 *Trans*-cleavage activity of different Cas12a orthologs using variants of Stem-loop #10T reporter. 317

D.6 Kinetic analysis of LbCas12a *trans*-cleavage activity. 318

D.7 Kinetic analysis of LbCas12a *trans*-cleavage activity on Stem-loop #10T (left) and Linear ss-DNA (right) reporters. 319

D.8 Michaelis-Menten analysis. 320

D.9 *Trans*-cleavage activity of Cas12a on Variant-ssDNA and Stem-loop #10T reporters. 321

D.10 The binding free energy for the stem loop and linear ssDNA binding Cas12a. 322

D.11 Contact analyses of the *trans*-cleavage substrates (i.e., stem loop and linear DNA) with the Nuc domain along the MD simulations. 323

D.12 *Trans*-cleavage reaction profile over time to evaluate the specificity of LbCas12a in the presence of wilt type and mutant dsDNA target. 324

D.13 Specificity test on single-nucleotide mismatch mutations (from MM1 to MM20) and double- nucleotide mismatch mutations (from MM1-2 to MM19-20) on synthetic SARS-CoV-2 dsDNA activator at different concentrations (10 nM, 1 nM, 200 pM). 324

D.14 Time-dependent *trans*-cleavage activity of LbCas12a activated by native ds-DNA (PM) and double-point mutations ds-DNA (MM16-17) activators. 325

E.1 FEP Schematic Diagram. 332

E.2 Nomenclature of sequences used in this study. 342

E.3 Free energy profiles obtained through Umbrella Sampling (US) simulations of the wt-TadA and TadA8e. 343

E.4 Design of ABE FRET constructs and biochemical and biophysical characterization of dye-labelled ABE0.1 and ABE8e. 345

E.5 Gain/loss of interaction strength at the Tada–Cas9–DNA triad of interfaces between ABE8e and ABE0.1. 347

E.6 Gain/loss of interaction strength at the Tada–Cas9–DNA triad of interfaces between ABE8e and ABE7.10. 348

E.7 Gain/loss of interaction strength at the Tada–Cas9–DNA triad of interfaces between ABE8e and ABE8e including a heterodimeric TadA8e (ABE8e hetero). 349

E.8 Critical interactions of the TadA docking unit in ABE8e and ABE7.10. . 350
E.9 Design of constructs for evaluation of TadA8e docking domain's impact on DNA adenine deamination efficiency. 351
E.10 The impact of TadA8e docking domain on DNA adenine deamination efficiency. 353
E.11 Convergence of the relative free energy changes obtained through alchemical free energy perturbation (FEP) simulations. 354
E.12 Denaturing polyacrylamide gels representing single-turnover kinetics of ABE7.10 variants. 354
E.13 Flexibility of TadA7.10 and TadA8e from molecular dynamics simulations of ABE7.10 and ABE8e, respectively. 355
E.14 Intra-domain interactions established by the 111-119 loop in ABE7.10 and ABE8e. 356
E.15 Inter-domain interactions established by the 111-119 loop in ABE7.10 and ABE8e. 357
E.16 Biophysical and biochemical characterization of dye-labelled ABE8e (T133C/N119D/N122H). 358

List of Tables

4.1 Cryo-EM Data Collection and Model Statistics 94

5.1 FRET-based DNA reporters 141

5.2 Guide crRNA 143

5.3 Single- and double-stranded DNA activators 144

B.1 Description of the simulated complexes. 257

B.2 Details for the calculation of bending angles using vectors. 259

B.3 Sequences of crRNA guide and target DNA substrates. 259

B.4 Sequences of DNA oligos, synthetic gene and gene fragments (gBlocks). 260

B.5 DNA Cleavage Assay Time Points Data. 270

C.1 Key Resources Table 294

C.2 Sequences of RNA (Synthego) and DNA (IDT) used for sample preparation and biochemistry reactions. 298

C.3 List of observed cleavage rates measured at 37 °C. 300

C.4 Competitor Cleavage Rates 301

D.1 PAGE analysis of LbCas12a *trans*-cleavage activity. 310

D.2 Measured K_m, k_{cat}, and k_{cat}/K_m. 311

D.3 Comparison of enzyme kinetic parameters. 312

E.1 ABE Variants. 339

Chapter I

Introduction

1.1 Overview of CRISPR-Cas systems

Genome editing is a sophisticated form of genetic engineering that involves inserting, deleting, or replacing DNA within the genome of a living organism. This groundbreaking technology has transformed numerous fields, including biomedicine, biotechnology, and synthetic biology.[3–5] At the heart of genome editing is the creation of double-strand breaks (DSBs) at specific genomic loci, a mechanism that has been historically achieved using tools like TALENs (transcription activator-like effector nucleases) and zinc finger nucleases (ZFNs).[6,7] However, the application of these earlier methods has been limited due to their complex reprogramming requirements for targeting specific DNA sequences. 2012 marked the advent of CRISPR-Cas (clustered regularly interspaced short palindromic repeats, CRISPR-associated proteins) technology that has revolutionized the field of precise manipulation of nucleic acids.[8–10] Bacteria and other prokaryotes wield

a sophisticated array of CRISPR-Cas systems as integral components of their adaptive immune response against foreign nucleic acid incursions.[4,11,12] Through relentless advancements in both basic and applied sciences, CRISPR-Cas9 has emerged as the preeminent "gold standard" for genome editing applications, catalyzing innovations across various disciplines within the life sciences and beyond.[8,11,13] Its versatility has transcended traditional boundaries, extending its reach into diverse fields such as plant biotechnology, biofuel production, and beyond, solidifying CRISPR-Cas9's reputation as the quintessential "genetic scissor" that shapes nature's gene pool.[14] Furthermore, CRISPR-Cas12a stands as another transformative biomolecular complex within the CRISPR-Cas technology landscape.[1,2,15,16] Not only does it serve as a robust molecular scissor, but it also emerges as an exceptionally rapid, ultrasensitive and reliable tool for nucleic acid detection.[2,17–20]

In the realm of CRISPR-Cas systems, the intricate interplay between small CRISPR RNA (crRNA) molecules and a constellation of Cas proteins results in the formation of crRNA ribonucleoprotein (crRNP) effector complexes, which are crucial for the destruction of invasive nucleic acids.[8,16,21] Encoded by the cas operon and the CRISPR array within CRISPR loci, Cas proteins and crRNAs work in concert in a finely tuned process. The CRISPR array comprises short repetitive sequences known as direct repeats, interspersed with non-repetitive sequences called spacers, which are derived from foreign genetic elements.[22–27] This structure highlights the capability of bacteria and archaea to acquire and integrate genetic elements into their own genomes, illustrating how interactions between prokaryotic genomes and the environment drive their evolution.[28] The CRISPR array undergoes transcription to produce precursor crRNAs, which

then mature into functional crRNAs. This maturation process is typically mediated by a protein complex involving trans-activating CRISPR RNA (tracrRNA) and RNase III.[29,30] However, Cas12a (formerly known as Cpf1) processes its own pre-crRNA into mature crRNAs without the need for tracrRNA, distinguishing itself as a unique effector protein with both endoribonuclease and endonuclease activities.[31] The repeat regions generally consist of identical sequences, facilitating recognition by effector Cas proteins in a sequence-specific manner. In contrast, the spacers exhibit a diverse array of sequences that are recognized in a non-sequence-specific manner by effector Cas proteins. Leveraging the Watson-Crick base pairing, the spacer region primes the Cas protein for its nucleic acid target. Consequently, crRNAs act as guiding beacons, directing crRNP effectors to their specific targets for degradation.

In the intricate choreography of CRISPR-Cas systems, the recognition of targets by most crRNP effectors necessitates the presence of a short nucleotide sequence adjacent to the target, known as the protospacer adjacent motif (PAM).[32] These systems are categorized into two classes, further subdivided into six types and 19 subtypes.[33] Within the class 1 system (types I, III, and IV), crRNP effectors operate in tandem with multiple Cas proteins,[34,35] while in the class 2 system (types II, V, and VI), a single Cas protein takes the helm.[8,9,33] Structural investigations have unveiled a rich tapestry of diverse and sophisticated mechanisms underlying target degradation by crRNP effector nucleases across the various types of CRISPR-Cas systems. At the molecular level, endonucleases Cas9 (type II) and Cas12a (type V) represent class 2 RNA-guided enzymes that exploit the sequence specificity of guide RNA to precisely target and cleave complementary DNA sequences.[8,15,36] Upon identification of a short protospacer adjacent motif (PAM),

the endonuclease binds to the DNA, establishing base-pairing between the guide RNA and the target strand (TS), while displacing the non-target strand (NTS).[37,38] This pivotal discovery underscores the remarkable capacity of these Cas endonucleases to be reprogrammed with a single guide RNA, enabling recognition of virtually any DNA sequence adjacent to the PAM site, thereby catalyzing the advent of advanced genome-editing technologies.

1.2 CRISPR-Cas12a: A extraordinary genome-editing and nucleic-acid detection tool.

1.2.1 Structural Details

Crystallographic and cryo-EM data on Cas12a have elucidated its bilobed architecture, akin to Cas9 and several other CRISPR-Cas complexes.[36] Cas12a exhibits distinct functional domains and represent type V CRISPR-Cas systems.[39] It comprises a recognition (REC) lobe and a nuclease (NUC) lobe linked by wedge domain (WED) and bridge helix (BH).[1,2,16,40] The WED-III subdomain contains the RNase site and is responsible for processing and maturation of its own cRRNA.[31] The REC lobe encompasses two alpha-helical domains (REC1 and REC2), which in juxtaposition with the NUC lobe forms a positively charged central channel for the accommodation of the crRNA-target DNA heteroduplex.[41] In silico analyses of the REC lobe showed no detectable homologs outside of the Cas12a family and thus appear to have adopted novel structural folds.[41] The NUC lobe houses the RuvC and Nuc domains, flanked by the PAM-interacting

(PI) domain, which recognizes the PAM sequence in the DNA substrate. As opposed to Cas9's two separate catalytic domains, HNH and RuvC,[13] Cas12a employs a single active site within the RuvC domain for cleaving both DNA strands.[1,2] This domain is consisted of three discontinuous parts (RuvC I-III) and the DNase site is located at the juxtaposition of the RuvC and the Nuc domains. This characteristic nuclease domain (i.e., RuvC) has been shown to be related to IS605 family transposon encoded TnpB proteins.[39] Intriguingly, Cas12a exhibits a sequential cleavage pattern, initiating with the NTS followed by the TS.[1,2,42]

1.2.2 Guide RNA (crRNA) recognition and processing

Cas12a possesses a ribonuclease site in the WED domain, which enables autonomous processing of pre-crRNA to mature crRNAs, without requiring the trans-activating crRNA (tracrRNA) necessary for Cas9.[31] The pre-crRNA displays a 19-nucleotide repeat fragment and a 50-nucleotide long repeat-spacer fragment. The segments derived from repeats within the pre-crRNA transcript are intricately engaged by Cas12a in a pseudoknot conformation.[2,41,43] *In vitro* cleavage assays showed that pre-crRNA processing is highly repeat-dependent, wherein the pseudoknot repeat region is specifically recognized by the WED, BH, and RuvC domains of Cas12a. This underscores the role of the pseudoknot structure, specifically its length and sequence for the binding and autonomous pre-crRNA processing by Cas12a.[15] Cryo-EM studies showed a hydrated Mg^{2+} ion maintaining an octahedral shape at the center of the pseudoknot structure.[31,43] Bioinformatic analyses of this phenomenon inferred type V as a connecting link between class 1 and class 2 CRISPR systems.[15]

Notably, the maturation of crRNA in Cas12a is finely orchestrated steps within the WED domain by the highly conserved KKH residues (lysine–lysine–histidine), facilitating the 5'-end processing of the pre-crRNA via acid–base catalysis.[2,15,31] This results into 19 repeat-derived nucleotides followed by approximately 24 spacer-derived nucleotides. In mature crRNA, residues from both the WED and REC1 domains form interactions with the phosphate and ribose moieties of the initial five nucleotides within the spacer-derived segment.[2] Consequently, this region, often termed the 'seed segment', assumes a preordered A-form helical conformation, with its bases exposed to the solvent and primed for binding with target DNA. Cas12a has been shown to be catalytically insert prior to the association with the pre-crRNA, devoid of any nuclease activity.[15] This phenomenon suggests that the binding of the crRNA pseudoknot induces conformational alterations across multiple domains of the Cas12a protein. It not only leads to independent maturation of the pre-crRNA but also primes it for subsequent target DNA binding and cleavage.

1.2.3 PAM Recognition and Target DNA Binding

The process of Cas12a–crRNA complex binding to target DNA commences with the recognition of a 5'-TT(T)NV-3' protospacer adjacent motif (PAM) in the NTS.[15,16] This recognition mechanism allows CRISPR-Cas systems in bacteria to distinguish between their own genomic DNA and foreign nucleic acids.[44] Notably, Cas12a effector proteins exhibit some flexibility, accommodating "suboptimal" PAMs containing cytosine but albeit with reduced efficiency.[45,46] Interestingly, Cas12a can bind away from the target sequence and diffuse bidirectionally in a random manner until it locates its target.[42,47]

Rapid equilibrium PAM binding results in local DNA melting which is followed by the interrogation of the target DNA by the crRNA.[2,48] The WEDII-III, REC1 and PAM-interacting domains are responsible for PAM recognition and initiating the hybridization of the 20-base pairs of complementary target DNA with the crRNA, a phenomenon regarded as R-loop formation.[49] PAM recognition is integral to Cas12a as it leads to ATP-independent target DNA strand separation and subsequent R-loop formation.[49,50] A conformational change is induced in the target DNA backbone by the insertion of the lysine helix from the conserved loop-lysine helix-loop (LKL region) of the PI domain, containing three conserved lysines (K667, K671, K677 in FnCas12a). A critical proline in the LKL region facilitates the LKL insertion and unzipping of the target DNA, also supported by mutational studies.[49] This kinking event prompts unwinding of the Watson-Crick interaction in the TS–NTS duplex downstream of the PAM region, facilitating the TS bases to reorient towards the pre-ordered seed segment of the crRNA guide. Binary complex of Cas12a revealed how the first five nucleotides of the crRNA expose and pre-order in A-form-like geometry, indicating the presence of a 5-bp seed.[2] This state of Cas12a is characterized by the outward shift of the REC1 domain and the RuvC-loop region (F1049 to Q1058 in AsCas12a and F1061 to Q1070 in FnCas12a) stabilizing the 5-bp intermediate, allowing progression of the R-loop formation.[51]

Subsequent base pairing occurs between the crRNA and the TS, beginning at the seed segment and extending towards the 3'-end of the crRNA, the accommodation of which is also aided by the dynamics of REC1 and dramatic conformational flexibility of the REC2 domain.[51] The BH maintain dynamic contacts between the REC2, the crRNA-TS heteroduplex and the NTS during this process.[51,52] During the formation

of the final few base pairs in the R-loop, all the domains dock stably and establish extensive contacts with the RNA-DNA heteroduplex.[51] As the crRNA–TS hybridization progresses, the NTS strand is displaced and is conducted toward the nuclease site in the RuvC domain, a process supported by residues within the PI domain and the positively charged "NTS-binding groove" and the "lid" region within the RuvC domain.[2,16,49,51] Structural studies of the cryo-EM intermediates revealed three regions as checkpoints to oversee the hybridization between the crRNA and the target DNA. They are: "REC Linker" connecting the REC1 and REC2 domains, "the lid" (L1008 - K1021) in the RuvC domain that obscures the active site in the apo structure and "the finger", an α-helix contained in the REC1 domain. While the REC linker and the finger exhibit notable dynamics to allow R-loop formation, the lid undergoes major conformational changes transitioning from a loop to an α-helix to detach from the active site and aid in the accommodation of the NTS at the active site.[1,51]

The stability of this crRNA-TS heteroduplex is further reinforced by the BH and conserved aromatic residues after approximately 20 base pairs, ultimately forming a robust "R-loop" configuration.[51–53] Single-molecule and biochemical assays have revealed a conserved aromatic residue (W355 in LbCas12a) in the REC2 domain that stacks against the end of the R-loop prevents R-loop extension during downstream DNA breathing. Mutation of this residue to alanine resulted in more stable R-loop formation, longer breathing states, and faster TS cleavage in the presence of a 3'-overhang in the crRNA complementary to the TS sequence.[53].

1.2.4 Allosteric Activation of Cas12a

Cas12a performs cleavage of both the target DNA strands (i.e., TS and NTS) by the catalytic DED (Asp-Glu-Asp) triad in the RuvC domain.[2,16] However, Cas12a maintains the catalytically inactive state even after binding with the crRNA (binary state) as a preventive measure against premature DNA cleavage. This conformation is characterized distinctly by: (i) close proximity of the REC2 domain to the catalytic RuvC and Nuc domains, and (ii) the loop form of the "lid" region in the RuvC domain positioned above the catalytic site, effectively blocking access to it. Consistent observations have been made by different structural studies, confirming that this closed conformation ensures inhibition of DNA cleavage until the complex binds to its DNA target.[1,43,54] These Cas12a structures necessitates notable REC lobe rearrangement alongside that of the lid to expose the RuvC active site for Cas12a to be catalytically competent.

Upon initiation of R-loop formation, triggered by the hybridization of crRNA with the target strand (TS), substantial conformational alterations occur within the Cas12a–crRNA complex.[1,16] Notably, the REC1 domain rotates towards the crRNA–TS duplex, while REC2 undergoes a rotational movement away from the nuclease (NUC) lobe. This repositioning of REC2 is crucial for averting clashes between the growing crRNA–TS heteroduplex and the REC2 domain. Concurrently, the lid region transitions from a closed to an open conformation, stabilizing the catalytic site upon DNA binding. While the lid region in the RuvC domain interacts with crRNA nucleotides 8–11 to unblock the RuvC active site, the finger region interacts with crRNA nucleotides 15–17, both playing pivotal role in the catalytic activation. Multiple studies have affirmed that a minimum of 17 base pairs between the crRNA and the TS are necessary for Cas12a to

attain its cleavage competent state.[48,55] These interactions likely facilitate the movement of REC2 away from the RuvC domain, thereby exposing the catalytic site to the solvent. Consequently, R-loop formation serves as an allosteric trigger for Cas12a activation.

Moreover, Cas12a exhibits versatility in target recognition, being activated not only by double-stranded DNA (dsDNA) but also by single-stranded DNA (ssDNA) targets complementary to the guide RNA. This suggests that while PAM binding is crucial for protospacer unwinding, it is dispensable for target DNA cleavage by Cas12a.[16,42] Collectively, these insights underscore the pivotal role of crRNA–TS hybridization in inducing conformational changes that transition Cas12a from its cleavage incompetent "closed" state to its cleavage competent "open" state.

1.2.5 *cis*- and *trans*-cleavages by Cas12a

Cas12a relies on a singular catalytic site within its RuvC domain to execute cleavage of both DNA strands.[2] Despite its nomenclature, the Nuc (nuclease) domain does not directly engage in nucleolytic activity; rather, it aids in guiding DNA to the catalytic site.[56] Within the RuvC domain, a positively charged 'NTS-binding groove' facilitates the guiding of the displaced single-stranded non-target strand (NTS) towards the catalytic center.[49] Additionally, steric hindrance purportedly prevents dsDNA from entering the RuvC catalytic center.[16] This structural arrangement implies a sequential cleavage mechanism where the NTS is cleaved prior to the target strand (TS). Thus, the cleavage of the NTS results from proper positioning of the DNA strand at the active site in the RuvC domain, rather than a requirement to initiate the cleavage mechanism for the target DNA. Notably, structural analyses of CRISPR-Cas12a complexes bound to cleaved NTS

reveal that the TS situates approximately 30 Å away from the active RuvC site, in close proximity to the REC2 domain.[56] Consequently, the delivery of such a spatially distant substrate to the RuvC site necessitates substantial conformational changes.

It is hypothesized that NTS cleavage facilitates local unwinding or "fraying" of the nicked DNA, with the Nuc domain likely playing a critical role in positioning the TS within the catalytic site.[57,58] This is supported by the observation that the Nuc domain have a more pronounced impact on TS cleavage than on NTS cleavage.[41] While the PAM-proximal end of the cleaved product maintains significant interactions with Cas12a, the PAM-distal end does not, leading to its release from the complex post-cleavage.[49] Despite this, the Cas12a complex remains in a cleavage-competent 'open' conformation as the crRNA and TS strands remain duplexed, resulting in further processing of the NTS and the production of heterogeneous cleavage products.

Beyond its primary cis-cleavage activity on target dsDNA, Cas12a retains an active conformation capable of facilitating non-specific *trans*-cleavage of incoming single-stranded DNA (ssDNA).[59] This unique property has been harnessed for nucleic acid detection applications.[17,18,60] Detailed biochemical and kinetic investigations have unveiled that Cas12a bound to an ssDNA target catalyzes trans-ssDNA cleavage at a rate of approximately 3 turnovers per second, exhibiting a catalytic efficiency (k_{cat}/K_M) of $5.0 \times 10^6\,\mathrm{s}^{-1}\,\mathrm{M}^{-1}$. Conversely, when bound to a dsDNA activator, Cas12a demonstrates a higher turnover rate of approximately 17 turnovers per second, with a k_{cat}/K_M of $1.7 \times 10^7\,\mathrm{s}^{-1}\,\mathrm{M}^{-1}$.[59] Notably, Cas12a also displays *trans*-cleavage activity towards ssRNA, indeed at a reduced efficiency compared to its ssDNA counterpart.[61] Intuitively,

such non-specific ssDNA cleavage poses a threat to the host cell and limits genome-editing applications by Cas12a. Upon in-depth investigations of the bacterial genomes encoding Cas12a and the crRNA, database search revealed that bacteria produces multiple copies (up to 68) of the crRNA. Considering that the transcription rates to be similar, the concentration of crRNA far exceeds that of Cas12a in the cell at any given time. Experimental data showed that indeed, sufficiently high concentration of crRNA can displace a cleaved R-loop from the Cas12a protein and reinstate the inactive state.[39] This inactivation is characterized by the titular lid closing on the active site and obscuring it from further catalytic activity. This not only prevents further ssDNA *trans*-cleavage activity, but also resets Cas12a for a new cycle of catalytic activity against new DNA targets.[1]

1.2.6 Off-target *cis*-cleavages in Cas12a

The specificity of DNA targeting mediated by the Cas12a–crRNA complex is remarkably high, characterized by a reversible formation of the R-loop until a stable configuration is achieved.[48] Typically, the establishment of a stable R-loop necessitates the presence of at least 17 base pairs between the crRNA and the target strand (TS), whereas stable R-loop formation may occur with as few as 9–10 base pairs.[55] Notably, R-loops formed by Cas9 exhibit a stability exceeding that of Cas12a by more than 10-fold. Cas12a requires at least $\sim$15 bp R-loop formation to activate cleavages mediated by the RuvC nuclease.[1,15,62,63] In the context of Cas12a, mismatches between the crRNA and the TS impede or entirely prevent R-loop formation, consequently accelerating the dissociation

rate of the target DNA. Recent structural studies have shown that mismatches in the 12-bp or shorter crRNA-TS heterduplex collapse at a faster rate than the R-loop formation, thereby acting as a defense mechanism against off-target effects.[42,51]

This relatively transient binding of Cas12a–crRNA complexes to their DNA targets has led to the proposition that Cas12a may serve as a more precise genome editing tool compared to Cas9.[55] Even a single mismatch between the crRNA and the TS can decelerate R-loop formation, yet typically permits the generation of double-stranded DNA breaks without notably affecting the kinetics of TS and non-target strand (NTS) cleavage.[48] Conversely, the presence of 2–4 mismatches between the crRNA and the TS alters the kinetics of TS and NTS cleavage, resulting predominantly in nicked DNA initially, with double-stranded DNA breaks occurring later. Furthermore, when more than four mismatches exist between the crRNA and the TS, only the non-target strand is nicked, while nicking of the target strand is prevented entirely.[64] Taken together, these findings highlight that although Cas12a can bind DNA even in the presence of up to four crRNA–TS mismatches, such DNA targets are predominantly nicked rather than cleaved. Thus, Cas12a is deemed as a more specific genome-editing tool than Cas9.[16,55]

1.3 CRISPR-Cas9: The golden genome-editing tool

1.3.1 A brief overview

CRISPR-Cas9, classified under class 2 and type II CRISPR-Cas families, functions as a dual RNA-guided DNA endonuclease responsible for targeted DNA degradation.

Utilizing a guide RNA (CRISPR-RNA; crRNA), Cas9 is directed to its specific target sequence.[8] Notably, investigations into the *Streptococcus pyogenes* Cas9 (SpCas9) system have unveiled the critical role of a secondary trans-encoded small RNA in crRNA maturation *in vivo*.[8,13] Biochemical studies in 2012 revealed that Cas9 forms an effector complex with both crRNA and a trans-activating crRNA (tracrRNA), cleaving double-stranded DNA targets complementary to the crRNA guide to generate blunt ends.

Cas9 harbors two endonuclease domains, namely HNH and RuvC. The HNH domain cleaves the DNA strand complementary to the crRNA guide, referred to as the target strand (TS), whereas the RuvC domain cleaves the displaced, single-stranded DNA strand, known as the non-target strand (NTS). Additionally, Cas9 requires a protospacer adjacent motif (PAM) located on the non-target strand for efficient dsDNA recognition. A significant milestone was achieved in 2012 with the development of a single-guide RNA (sgRNA) capable of directing Cas9 to its target for cleavage. Subsequent studies in 2013 demonstrated the robust activity of a simple two-component system comprising *S. pyogenes* Cas9 (SpCas9) and its sgRNA in mammalian cells. Today, the SpCas9–sgRNA system stands as a widely adopted genome engineering tool, facilitating precise genetic modifications across diverse biological contexts.

1.3.2 A comparative study of CRISPR-Cas9 with respect to CRISPR-Cas12a

Cas9 and Cas12a, despite sharing analogous functionalities, exhibit notable disparities in their evolutionary lineage and mechanistic behaviors. While *Streptococcus pyogenes* Cas9 targets a protospacer adjacent motif (PAM) sequence of 5'-NGG-3' directly

alongside the protospacer Non-Target Strand (NTS), Cas12a orthologs recognize a distinct 5'-TTTV-3' PAM situated upstream of the protospacer NTS. Intriguingly, Cas12a displays adaptability by proficiently targeting DNAs with "suboptimal" PAMs, such as 5'-TTV-3', 5'-TCTV-3', 5'-TCCV-3', and 5'-CCCV-3', albeit often with diminished efficiency. Mechanistically, Cas9 and Cas12a employ different strategies for DNA cleavage. Cas9 utilizes its dual catalytic domains, namely the HNH and RuvC domains, to cleave the Target Strand (TS) and NTS, respectively, typically resulting in blunt-end or 1-nucleotide overhang double-strand DNA breaks (DSBs) proximal to the PAM. In contrast, Cas12a relies on a single RuvC nuclease domain to cleave both strands of the target DNA, generating PAM-distal DSBs. The resulting product contains larger 5' overhangs ranging from 5 to 7 nucleotides, rendering it more conducive for genome editing applications. Despite their initial cleavage at a single site, both Cas9 and Cas12a enzymes facilitate trimming of the cleaved ends, yielding a variety of cleavage products in vitro. Notably, Cas12a exhibits two distinct modes of DNA cleavage: besides its single turnover crRNA-guided cis-cleavage of dsDNA targets, it also possesses multi-turnover sequence-independent cleavage capability, enabling trans-cleavage of incoming single-stranded DNA upon hybridization of the crRNA to a complementary target DNA strand. Moreover , target ssDNA hybridization with the crRNA alone is sufficient to allosterically activate Cas12a, suggesting that PAM binding with the PI domain is necessary for dsDNA unwinding. In contrast, Cas9 is demonstrated to be allosterically activated only by dsDNA with the binding of the PI domain on the PAM region of the NTS. Unlike Cas9, which necessitates 9 bp for stable binding and $\sim$16 bp for cleavage, Cas12a requires an $\sim$17-bp sequence match for both stable binding and cleavage.

Overall, Cas12a exhibits superior discrimination against both PAM-proximal and PAM-distal mismatches for stable binding. It has been conjectured that owing to the greater degree of DNA bending induced in Cas9, it harbors a higher tolerance for binding and cleaving mismatched sequences, i.e., off-target cleavages, than Cas12a. Furthermore , the binding specificity of engineered Cas9s (eCas9 and Cas9-HF1) is still markedly lower than that of Cas12a, suggesting that Cas12a holds potential as a superior alternative to all current Cas9 variants.

1.4 Therapeutic Applications of CRISPR-Cas Systems

Cas9 acts primarily as genome-editing tool and can be used as a navigating tool for other functions such as base editors and prime editors by conjugating them with the enzyme counterpart. CRISPR-Cas12a as a molecular tool serves not only as a genome-editing tool but also as a nucleic acid detection tool. As a gene editor, it can render double stranded DNA breaks at specific genomic loci and allowing the cellular DNA repair machinery by correcting the damage induced by the CRISPR-Cas system with the correct nucleotide inserts to maintain the genomic integrity. Two major pathways responsible for repairing DNA double-strand breaks (DSBs) are Homology-Directed Repair (HDR) and Non-Homologous End Joining (NHEJ). NHEJ is a more flexible but error-prone repair pathway that directly ligates the broken DNA ends without the need for a homologous template. HDR is an accurate repair mechanism that uses a homologous sequence as a template to repair the break, ensuring the original sequence is precisely restored. HDR can thus be used with exogenous donor DNA template that is designed

to contain the desired genetic change flanked by homology arms. These arms are sequences that are homologous to the regions immediately upstream and downstream of the DSB. The homologous sequences on the donor template align with the corresponding sequences on either side of the break, facilitating precise repair and integration of the desired genetic change. This is particularly useful for gene editing applications, such as correcting mutations, inserting genes, or making specific nucleotide changes.

CRISPR-Cas9 has been at the helm of genome editing owing to its facile reprogrammability and has drastically transformed the landscape of genome-editing technologies. It has launched unprecedented applications into plethora of different fields, starting from genetic therapy, pharmaceuticals, food industry and plant biotechnology. However, it suffers from several limitations, especially off-target effects and difficulty of delivering them into the cell. Cas12a has surfaced as a better alternative owing to it high specificity in cells[65,66] and staggered cleavage products.[15,48] Gene repression studies comparing the performance between catalytically dead *S. pyogenes* Cas9 (SpdCas9) and *Eubacterium eligens* Cas12a (EedCas12a) have unveiled better performance by the latter. Furthermore, the pre-crRNA processing activity of Cas12a makes it a more suitable molecular tool for genome-editing purposes. Animal model[67]

1.5 Diagnostic Applications of CRISPR-Cas12a

CRISPR-Cas12a remains active post its target DNA cleavage, and can cut any incoming single stranded DNA and RNA, regarded as *trans*-cleavage.[59] This property has been ingeniously repurposed for nucleic acid detection, especially for diseases with a genetic

marker.[68,69] In 2020, when the world was wrecked by the pandemic caused by SARS-CoV-2, the *trans*-cleavage property of Cas12a was harnessed to develop the SARS-CoV-2 DNA Endonuclease-Targeted CRISPR Trans Reporter (DETECTR) technology.[17] This approach involves rapid (¡40 min), easy-to-implement and ultrasensitive biosensors using lateral flow assays with 95% positive predictive agreement and 100% negative predictive agreement. While DNA samples can be readily amplified, This assay performs simultaneous reverse transcription and isothermal amplification using loop-mediated amplification (RT–LAMP) for RNA extracts. Next, appropriate guide/crRNA against the genetic marker guides the Cas12a endonuclease to render targeted DNA cleavage. Finally, single-stranded reporter DNA tagged with a fluorophore and quencher can be supplied to be cleaved by virtue of the indiscriminate *trans*-cleavage property of Cas12a. This emits fluorescence to confirm the detection of the gene. The Cas12a fluorescent signal could be detected in $<$1 min and a visual signal by lateral flow within 5 min, thereby allowing the overall process to take $<$40 min and achieves accuracy comparable to the current gold-standard qRT-PCR.[17] Another group of researchers developed an Autocatalytic Cas12a Circular DNA Amplification Reaction (AutoCAR) system that has been demonstrated to show 1 aM ($<$1 copy/μL) sensitivity toward DNA detection without the requirement of any additional amplification step.[18] This autocatalysis approach also involves DNA nanostructures modified with fluorophore-quencher pairs and can detect a genetic marker in $<$15 min under room temperature. The detection range is tunable from 3 to 11 orders of magnitude and this technology has been successfully implemented in SNP mutations in circulating tumor DNA from blood plasma, genomic DNA (*H. pylori*) and RNA (SARS-CoV-2) without reverse transcription.

1.6 Overarching Questions

Cas12a stands as a versatile endonuclease with a vast array of applications. However, this versatility and the aforementioned premise prompts several overarching questions that necessitate a comprehensive investigation into this fascinating biomolecular complex. First, the activation of the Cas12a protein involves a significant conformational shift at the molecular level, which is crucial for DNA accommodation and cleavage. However, the mechanism underlying this conformational transition remains largely enigmatic. This topic represents a focal point of intense research, investigated through structural, kinetic, and single-molecule experiments. These studies suggest that the conformational transitions occur rapidly, potentially within microseconds or sub-milliseconds, and exhibit significant plasticity, particularly within the REC2 and Nuc domains. Despite these advancements, the precise manner in which these conformational changes in these domains trigger DNA cleavage or facilitate the exchange of the target strand (TS) and non-target strand (NTS) within the RuvC active site to achieve double-stranded DNA (dsDNA) breakage remains unclear.

Second, cleavage of both the strands of the target DNA by the RuvC domain alludes to a conundrum of how a single nuclease domain could enable cleavage of both DNA strands in Cas12a. Indeed, while the NTS readily accommodates within the RuvC catalytic cleft, the TS stays spatially distant (30 Å) from the catalytic pocket, close to the REC lobe. Experimental evidence suggests that the TS undergoes cleavage following the cleavage of the NTS.). Kinetic studies have shown that the clevage of the TS is approximately ten times slower than the NTS. This observation implies a conformational change after the NTS cleavage, enabling the TS to be accommodated within the RuvC

catalytic pocket. However, the molecular intricacies of these fascinating biophysical processes remain incompletely understood.

Third, the Nuc domain's emergence in Cas12a literature initially sparked controversy, with early speculation suggesting its role in cleaving the target strand (TS), akin to HNH in Cas9, which was later debunked. Recent studies have indicated that the Nuc domain instead contributes to binding both the non-target strand (NTS) and the TS DNA. Moreover, the positioning of the Nuc domain near the RuvC domain at the active site introduces steric hindrance, preventing the binding of non-specific double-stranded DNA during trans-cleavage and thereby safeguarding target cells from abortive infection. Understanding the dynamics of the Nuc domain is crucial for elucidating its role in regulating the binding of trans-cleavage substrates, particularly in applications such as nucleic acid detection in DETECTR technology. Moreover, the deliberate engineering of alternative DNA reporters exhibiting heightened affinity for Cas12a proteins, thereby bolstering sensitivity and trans-cleavage rates, represents an area of research that has received inadequate attention. Notably, short, single-stranded DNA probes have emerged as superior candidates for FRET-based reporting of trans-cleavage activity when juxtaposed with their double-stranded counterparts. Nonetheless, the impact of distinct DNA secondary structures, such as hairpins or triplexes, on the performance of these reporters remains under-explored.

Chapter II

Molecular Dynamics Reveal a DNA-Induced Dynamic Switch Triggering Activation of CRISPR-Cas12a

Reprinted with permission from Saha A. *et al.*

2.1 Abstract

CRISPR–Cas12a is a genome-editing system, recently also harnessed for nucleic acid detection, which is promising for the diagnosis of the SARS-CoV-2 coronavirus through the DETECTR technology. Here, a collective ensemble of multi-microsecond molecular dynamics characterizes the key dynamic determinants allowing nucleic acid processing

in CRISPR-Cas12a. We show that DNA binding induces a switch in the conformational dynamics of Cas12a, which results in the activation of the peripheral REC2 and Nuc domains to enable cleavage of nucleic acids. The simulations reveal that large-amplitude motions of the Nuc domain could favor the conformational activation of the system toward DNA cleavages. In this process, the REC lobe plays a critical role. Accordingly, the joint dynamics of REC and Nuc shows the tendency to prime the conformational transition of the DNA target strand toward the catalytic site. Most notably, the highly coupled dynamics of the REC2 region and Nuc domain suggest that REC2 could act as a regulator of the Nuc function, similar to what was observed previously for the HNH domain in the CRISPR-associated nuclease Cas9. These mutual domain dynamics could be critical for the non-specific binding of DNA and thereby for the underlying mechanistic functioning of the DETECTR technology. Considering that REC is a key determinant in the system's specificity, our findings provide a rational basis for future biophysical studies aimed at characterizing its function in CRISPR-Cas12a. Overall, our outcomes advance our mechanistic understanding of CRISPR-Cas12a and provide a ground for novel engineering efforts to improve genome editing and viral detection.

2.2 Introduction

CRISPR–Cas (clustered regularly interspaced short palindromic repeats and CRISPR–associated proteins) are a part of bacterial immune system that confer protection against invading viruses. In 2012, the discovery that the CRISPR–associated protein Cas9 is an RNA-programmable endonuclease[8] enabled precise manipulation of nucleic acids,

launching an unprecedented genome editing revolution.[3] Recently, a novel CRISPR protein, Cas12a[15] emerged as a promising tool for innovative applications of the CRISPR technology, such as nucleic acid detection.[59] The CRISPR-Cas12a system is the basis of the DETECTR technology,[17] which allows rapid detection of viruses including SARS-CoV-2, which is spreading across multiple countries.

At the molecular level, the CRISPR-Cas9 and CRISPR-Cas12a systems confer precise double-stranded DNA (dsDNA) breaks by using CRISPR RNAs (crRNA) as guide for molecular recognition of substrate DNA.[3] Thanks to a short protospacer adjacent motif (PAM) sequence in the viral DNA, these systems can be programmed to recognize any DNA sequence of interest, therefore enabling genome editing.[44] Despite sharing functional similarities with CRISPR-Cas9, Cas12a has a distinct evolutionary history and intriguing mechanistic properties.[39,70,71] Structural and biochemical studies of Cas12a have revealed a bilobed architecture, similar to what was observed in Cas9. It comprises a recognition (REC) lobe and a nuclease (NUC) lobe connected by a wedge domain (WED, Figure 2.1).[1,2,16,40,41,43,72] The REC lobe includes two alpha-helical domains (REC1 and REC2), which mediate nucleic acid binding. The NUC lobe consists the RuvC and Nuc domains, flanked by the PAM-interacting (PI) domain, which binds to the PAM sequence in the DNA. The guide RNA forms a heteroduplex with one DNA strand (the target strand TS), while the other non-target strand (NTS) is accommodated within a cleft formed by the RuvC and Nuc domains (Figure 2.1b,c). Unlike Cas9, which cleaves the TS and NTS using two specific catalytic domains, HNH and RuvC, respectively, Cas12a performs cleavage of both the DNA strands using a single active site located within the RuvC domain.[2] This has raised questions on the role of the

additional Nuc domain, which was initially thought to cleave the DNA TS, enacting the role of HNH in Cas9.[41] More recent studies have implied that the Nuc domain instead plays a role in NTS and DNA binding.[1,2,16] However, it is unclear how conformational changes of the Nuc domain would activate DNA cleavages or facilitate the exchange of the TS and NTS within the RuvC active site to achieve a dsDNA break.

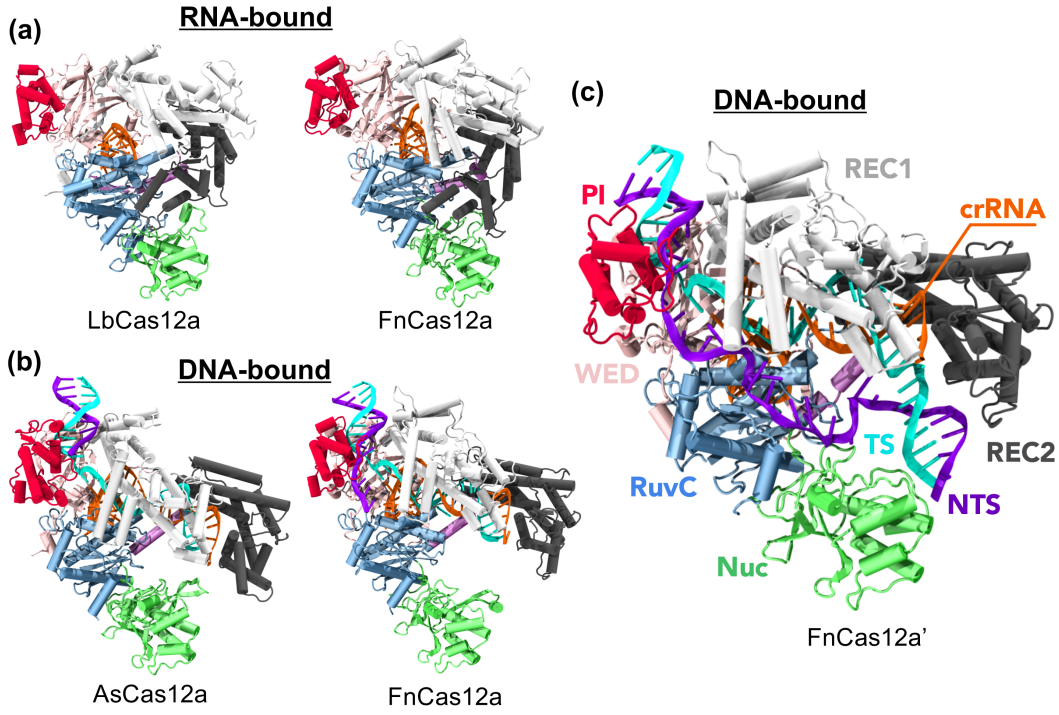


Figure 2.1: **Overview of CRISPR-Cas12a.** X-ray structures of Cas12a proteins in complexes with guide RNA **(a)**, or guide RNA and target DNA **(b, c)** across different species, *viz.* *Lachnospiraceae bacterium* Cas12a (LbCas12a),[43] *Acidaminococcus sp.* Cas12a (AsCas12a)[41,72] and *Francisella novicida* Cas12a (FnCas12a).[1,2,16,40] Cas12a proteins are shown in cartoon format, highlighting individual protein domains in different colors. The guide RNA (orange), the DNA target strand (TS, cyan), and the DNA non-target strand (NTS, violet) are shown as ribbons. The DNA-bound states include a cleaved NTS **(a)** and a complete NTS **(b)**.

Atomic resolution structures have captured Cas12a bound to guide RNA alone (for *Lachnospiraceae bacterium* LbCas12a10 and *Francisella novicida* FnCas12a13) and

to a guide RNA and a target dsDNA (for *Acidaminococcus sp.* AsCas12a12 and FnCas12a,[2] Figure 2.1). For the DNA-bound states, both AsCas12a and FnCas12a have been determined in complex with a cleaved NTS (Figure 2.1b),[1,40,41] while FnCas12a has also been obtained including a longer NTS that binds within RuvC and reconciles with the TS (Figure 2.1c, hereafter referred to as FnCas12a').[16] Collectively, these atomic-resolution structures offered intricate details about the Cas12a ribonucleoprotein complex, suggesting extensive conformational plasticity. However, it is unclear how this structural plasticity could contribute to the "open–to–closed" conformational change of the protein, which is thought to facilitate substrate DNA binding and subsequent cleavage.[1] In light of these experimental findings, investigating the protein structural plasticity at the atomic level through Molecular Dynamics (MD) is critical for understanding its biological functions and rational engineering of novel Cas12a-based tools for genome editing and viral nucleic acid detection.

Molecular simulations have previously contributed in the understanding of fundamental biophysical aspects of the CRISPR-Cas9 system,[73–78] revealing a striking plasticity of the ribonucleoprotein complex.[73–77] These studies provided predictive insights into the dynamic behavior of Cas9, many of which have been corroborated by single-molecule FRET experiments[79] and cryoEM,[80] providing also a framework for rational design of Cas9 for improved genome editing.[81,82] Here, we present the first multi-microsecond length MD study of the CRISPR-Cas12a, to characterize the conformational plasticity of the protein and its interplay with the nucleic acids. We collected an overall ensemble of $\sim$20 μs, carried out in multiple replicates, considering different states and across species. The simulations reveal a "switch" in the conformational

dynamics of Cas12a upon DNA binding that results in the activating motions of the peripheral REC2 and Nuc domains. In agreement with previous structural[2,40] and single-molecule experiments,[1] this increased mobility of REC2 and Nuc upon DNA binding could enable the conformational changes associated with DNA cleavage. The simulations also reveal an important role of the Nuc domain, whose large–amplitude motions could enable the conformational activation of the system for completing DNA cleavages. In this process, the REC lobe is critical in aiding the conformational dynamics of Nuc. Indeed, highly coupled dynamical motions of REC2 and Nuc suggest that REC2 could act as a regulator of the Nuc function, as previously observed for the HNH domain in CRISPR-Cas9.[75,79–81,83] Considering the key role of the REC lobe for the specificity of Cas9,[65,81,84] and as recently found in Cas12a,[85] our outcomes now call for future studies aimed at characterizing the functional role of REC and Nuc in CRISPR-Cas12a.

2.3 Results

2.3.1 Conformational Flexibility of Cas12a Bound to RNA and DNA

Building on the available structures of Cas12a (Figure 2.1), we collected ensembles of μs-length MD simulations. We examined LbCas12a, AsCas12a and FnCas12a to assess differences and similarities in the dynamics across various species and states (i.e., crRNA-bound *vs.* DNA-bound). For each model system of Cas12a, MD simulations were carried out in explicit solvent, obtaining multiple μs-length trajectories (i.e., 4 replicates of $\sim$1 μs, each) and reaching an overall sampling of $\sim$20 μs. The choice of

simulating multiple and independent μs-length trajectories was motivated by the need to achieve solid statistics for the purpose of our analysis, and by our previous theoretical investigations of the parent CRISPR-Cas9.[73–75] Those studies had shown that multiple ns–to–μs MD trajectories are critical for describing the inter-dependent dynamics of the protein domains and their interplay with the nucleic acids. Notably, in the present work, data analysis was performed in analogy to our early multi–μs MD investigations of CRISPR-Cas9.[73] This enabled a fair comparison of the dynamical properties between the two CRISPR-Cas systems.

To develop an initial understanding of the overall flexibility of the system when bound to RNA and DNA, we employed root mean square fluctuation (RMSF) analysis, which is a traditional mode of measuring protein flexibility (Supplementary Figure A.1). Additionally, to investigate whether the observed changes in fluctuations are maintained across the ns–to–μs runs, we computed a time-dependent RMSF (t-RMSF, Figure 2.2a). As a result, the t-RMSF reveals high fluctuations of the PI domain in the RNA-bound LbCas12a and FnCas12a (Figure 2.2a) complexes, which are conserved along the simulated runs. Upon DNA binding, the flexibility of the PI region is remarkably reduced for both AsCas12a and FnCas12a, as stabilized by the binding of DNA. Simultaneously, the flexibility of REC2 and Nuc shows a remarkable increase. The change in flexibility of the PI, REC2 and Nuc domains upon DNA binding is shown on the 3D structure of FnCas12a, indicating regions of high fluctuations using thicker tubes (Figure 2.2b). Interestingly, a high flexibility of REC2 and Nuc is preserved in the FnCas12a' complex, where a dsDNA locates in between REC2 and Nuc (Figure 2.2b, lower panel). This is notable because the binding of a dsDNA commonly stabilizes the surrounding protein

scaffold, as observed at the level of the PI domain. On the other hand, the RMSF of the RuvC domain, which is responsible for both NTS and TS cleavages, reveals low fluctuations (Supplementary Figure A.11). This is in agreement with high structural stability of this conserved domain,[86] also observed in MD simulations of CRISPR-Cas9.[73] The remaining protein domains do not show significant flexibility upon DNA binding. Analysis of the root mean square deviation (RMSD) of the protein Cα atoms reveals that the protein backbone reaches a similar stability in both RNA- and DNA-bound states (i.e., the RMSD reaches ∼4-5 Å, Figure S2). This indicates that upon DNA binding, the overall stability of the protein is preserved, but the flexibility of different protein domains changes, as shown by the analysis of the *t*-RMSF (Figure 2.2a).

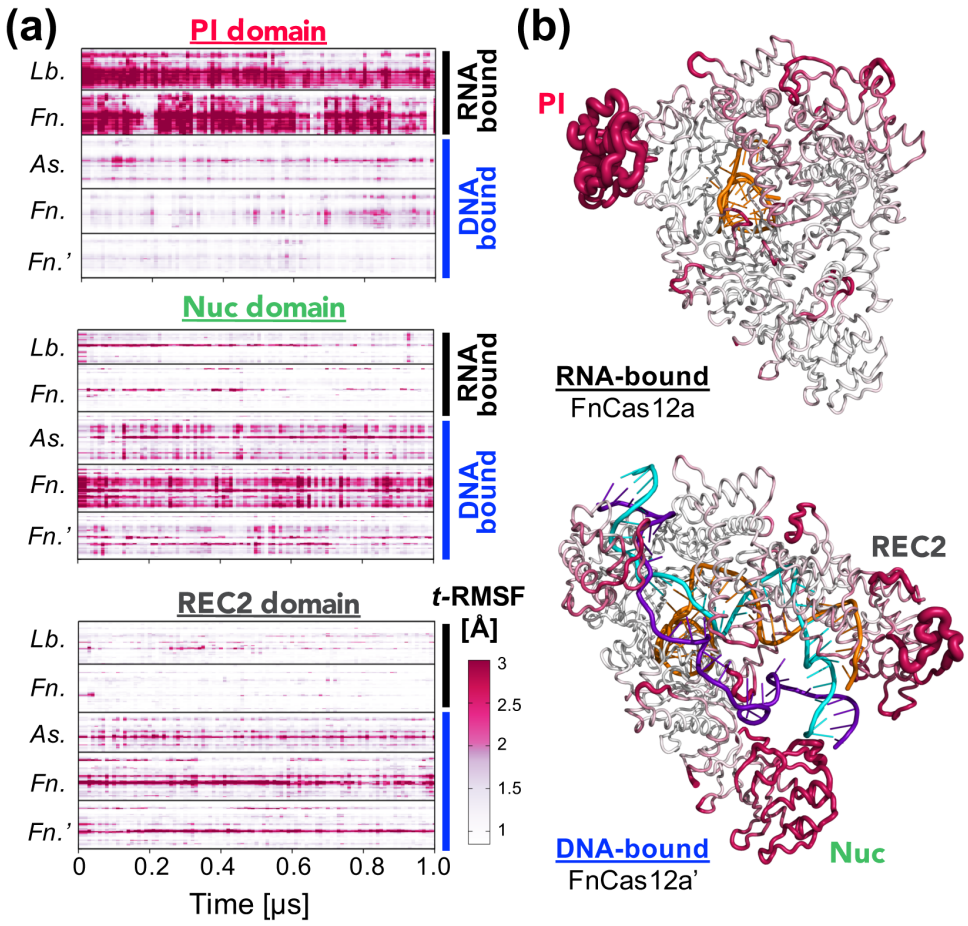


Figure 2.2: **Cas12a flexibility along the dynamics.** **(a)** Time-dependent Root Mean Square Fluctuations (*t*-RMSF, in Å), computed for the PI (top), Nuc (center), and REC2 (bottom) domains of the LbCas12a (*Lb.*), AsCas12a (*As.*), and FnCas12a (*Fn.*) systems. The RNA-bound and DNA-bound systems are indicated using black and blue bars, respectively. *t*-RMSF values are colored from white (low fluctuations) to magenta (high fluctuations), accordingly to the scale on the bottom right. **(b)** The averaged RMSF values are plotted on the 3D structures of the RNA-bound FnCas12a (top) and DNA-bound FnCas12a' (bottom), indicating protein regions of high fluctuations using thicker tubes (color-coded according to the *t*-RMSF scale). The RNA (orange) and the DNA TS (cyan) and NTS (violet) are shown as cartoon.

2.3.2 Large–amplitude Motions and Conformational Ensemble

Aiming to dissect the large–scale collective motions of the Cas12a-nucleic acid complexes and characterize its essential degrees of freedom, we performed Principal Component Analysis (PCA). With this analysis, the "essential dynamics"[87] of the system is

described along the first principal mode of motion (i.e., principal component 1, PC1), providing information on the large–amplitude motions of the complexes and, in turn, on their functional dynamics. PCA has been performed considering all the FnCas12a systems (i.e., the RNA-bound FnCas12a, DNA-bound FnCas12a and FnCas12a'). In detail, we combined the collected ensembles arising from the compared systems and subjected to RMS-fit to the same reference configuration, ensuring consistency of the eigenbasis and motions of the PCs (details are reported in the Methods section).

In Figure 2.3a, PC1 is plotted over the 3D structures of FnCas12a in the RNA- and DNA-bound states, where the arrows indicate the direction and relative amplitude of motions. In the RNA-bound state, the PI domain displays large–amplitude motions, which are directed toward the cleft that accommodates the PAM of the DNA. These motions agree well with previous structural analyses, suggesting that the inward movement of the PI domain would accompany PAM binding.[40,72] Upon binding of DNA, the PI domain remarkably reduces the amplitude of its motions, while REC2 and Nuc display substantially increased amplitude motions. The large–amplitude of the motions of REC2 and Nuc is preserved in all DNA-bound states, in the presence of a cleaved NTS (i.e., in the FnCas12a system) and in the FnCas12a' system, where a complete NTS rehybridizes with the TS to form a duplex in the vicinity of REC2 and Nuc (Figure 2.1c and 2.3a). As previously suggested by Stella and co-workers based on single-molecule FRET,[1] an increased flexibility of REC2 and Nuc upon binding of DNA could have a functional role. Indeed, the mobility of these regions could favor the exchange of the NTS and TS, as well as to attain the 5'-3' polarity of the TS required for cleavage within the RuvC active site.[1] In this respect, the catalytic RuvC domain displays short–amplitude

motions, in agreement with the high structural stability of this conserved protein domain.[73,86] Interestingly, REC1 shows motions of smaller amplitude, which arise from strong interactions with the RNA:DNA hybrid and reflect the function of this domain in anchoring the target DNA strand.[40]

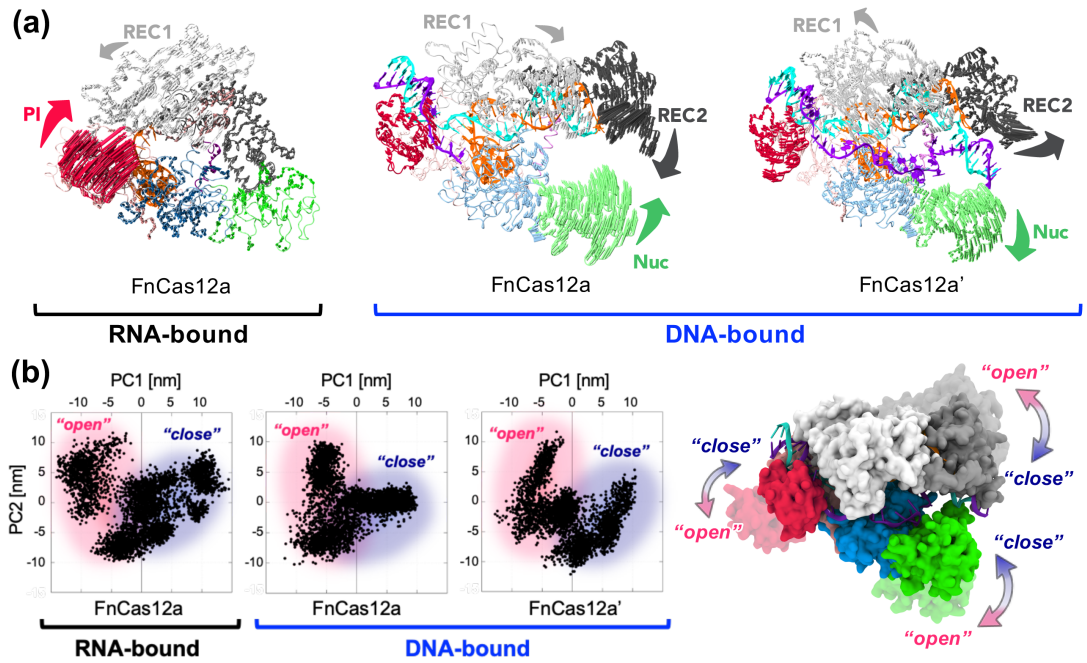


Figure 2.3: **Large-amplitude motions and conformational space adopted by CRISPR-Cas12a.** **(a)** "Essential dynamics", derived from the first principal component (PC1) of the individual protein domains of FnCas12a bound to RNA (left) and DNA (right), shown using arrows of sizes proportional to the amplitude of motions. For the DNA-bound states, the "essential dynamics" is shown for the presence of a cleaved NTS (FnCas12) and complete NTS (FnCas12a'). **(b)** Projections of the first and second principal components (PC1 *vs.* PC2) from MD simulations of the RNA- and DNA-bound FnCas12a systems. The PC1 *vs.* PC2 plots characterize the conformational space sampled by the FnCas12a into regions in which the protein is "open" (red cloud) and "closed" (blue cloud). A cartoon of FnCas12a indicating the "open-to-closed" conformational transition is shown on the right.

The direction of FnCas12a principal motions upon DNA binding is of particular interest, as differences are observed in the presence of a complete NTS, as opposed to cleaved NTS. In FnCas12a, in which the NTS is cleaved, REC2 and Nuc move

toward each other (Figure 2.3a), and REC1 also moves toward RuvC and Nuc. In the FnCas12a' system (including a complete NTS), REC1-2 and Nuc preserve opposite motions but directed away from each other. Nuc points its motions out of the protein framework, as also observed for REC1-2. To gain further insights on this observation, we performed volumetric analysis on the equilibrium trajectories (details are reported in the SI). We measured the volume of the cavity between the REC and NUC lobes, which forms the RuvC binding groove. This revealed a contraction of the groove in the FnCas12a system (Supplementary Figure A.6). On the other hand, in the FnCas12a' system, an expansion of the groove is observed. Considering that Cas12a cuts the NTS first,[1,2,42] the dynamical differences observed in the presence and absence of a complete NTS, suggests possible conformational rearrangements of Nuc and REC1-2 upon NTS cleavage, which could allow the subsequent processing of the TS.[1]

Overall, the direction of the motions of these domains (as opposite to each other) indicates the tendency toward the "opening" and "closure" of the protein to accommodate and cleave the nucleic acids.[1,40,72] This is a functional feature shared with CRISPR-Cas9, as observed in previous simulation studies and through structural analyses.[73,88,89] To characterize the conformational space adopted by the ribonucleoprotein, we plotted the first versus the second principal components (i.e., PC1 *vs.* PC2, Figure 2.3b). As a result, the PC1 *vs.* PC2 plots identify two states, which depict well the "open" and "closed" conformations of the protein (schematically drawn using a cartoon of FnCas12a in Figure 2.3b), observed in both the RNA-bound and DNA-bound forms of FnCas12a. Notably, a range of conformational states from "open" to "closed" was also observed through PCA of multi–μs MD runs on CRISPR-Cas9.[73,89] This indicates that

both ribonucleoprotein complexes have a general tendency toward an "open–to–closed" breathing to allow for nucleic acid association. In this respect, it is worth noting that DNA binding-induced conformational changes occur at remarkably slower rates than what is possible to simulate using classical MD.[90,91] Yet, our data show a good coverage of the conformational landscape (Figure 2.3b and Supplementary Figure A.7), implying that the PCA well-represented the large-scale dynamics of the system. It is also notable that in the DNA-bound states, the conformational space explored by the protein is slightly restricted as a result of global stabilization due to the binding of the DNA (FnCas12a) and of a complete NTS (FnCas12a'). This also agrees well with previous PCA of the CRISPR-Cas9 complex.[73] In summary, the Cas12a protein preserves a general tendency toward an "open–to–closed" conformational transition in both the RNA-bound and DNA-bound states. However, as described above, prior to DNA binding, large–amplitude motions and high flexibility are observed at the level of the PI domain. On the other hand, upon DNA binding, the largest motions shift to REC2 and Nuc.

Finally, it is noteworthy that the PCs arising from the LbCas12a, AsCas12a and all FnCas12a systems cannot be directly compared owing to the inconsistency in the eigenbasis. Acknowledging this fact, two independent PCA have been performed for the RNA-bound LbCas12a and DNA-bound AsCas12a systems to gain insights into their "essential dynamics". We observed that the RNA-bound LbCas12a displays large amplitude motions in the PI domain directed inward. On the other hand, AsCas12a bound to a cleaved NTS exhibits the largest motions in the REC2 and Nuc domains, moving toward each other (Supplementary Figure A.7). This is qualitatively consistent

with the switch in the "essential dynamics" observed in the FnCas12a systems upon DNA binding.

2.3.3 Coupled Motions of Protein Domains

To investigate the inter-dependent conformational dynamics among spatially distant protein domains, we performed dynamic correlation analysis. This analysis was performed and averaged over multiple MD trajectories using two different methods. We computed the traditional Pearson cross-correlation (CC_{ij}) coefficients, which measure the collinear correlation between two Cα atoms (i and j), determining whether they tend to move in lockstep (positive CC_{ij}) or show opposed motions (negative CC_{ij}). The CC_{ij} analysis only detects correlations that are collinear with each other, neglecting correlated motions occurring out of phase. Hence, we also employed a generalized correlation (GC_{ij}) scheme.[92] This measures the degree of correlation between Cα atoms based on their mutual information, providing a normalized measure of how much information on one atom's position is provided by that of another atom. The method, however, does not distinguish positive *vs.* negative correlations, neglecting the description of opposite atom's motions. Hence, when employed together, the CC_{ij} and GC_{ij} schemes that are powerful in describing the inter-dependent dynamics of proteins.

The CC_{ij} matrix (i.e., a two-by-two plot of the Cα CC_{ij} coefficients) of FnCas12a shows a conserved pattern of correlated/anticorrelated motions in both RNA- and DNA-bound states (Figure 2.4, upper triangles), which are also found in the other Cas12a systems (Supplementary Figures A.8 - A.12). The REC lobe (i.e., REC1-2) preserves anticorrelated motions with the NUC lobe (including RuvC, Nuc, PI and WED). This

indicates the tendency of REC to move in an opposite way with respect to NUC, thereby favoring the "open–to–closed" conformational transition underlying nucleic acid binding.[1,40,72] The GC_{ij} matrix, which goes beyond the reach of a Pearson-like CC_{ij} analysis, captures the overall dependencies of the protein motions (Figure 2.4, lower triangles). In the RNA-bound state of FnCas12a (Figure 2.4a), coupled motions are only detected among REC1-2 and PI. On the other hand, in the DNA-bound FnCas12a (Figure 2.4b) correlated motions of REC2 and Nuc become prominent, while REC1 also displays correlations with Nuc (although at a lower extent). In the presence of a complete NTS (i.e., in the FnCas12a', Figure 2.4c), the overall system's GC_{ij} becomes more intense, preserving a high degree of coupling between REC1-2 and Nuc. We note that intense correlations between REC2 and Nuc are found in all simulated replicas of each DNA-bound system (Supplementary Figures A.8 - A.10). This indicates a shift in the correlated motions from one region of the GC_{ij} matrix (corresponding to REC1-2 and PI, in the RNA-bound state) to the REC1-2/Nuc domains upon DNA binding, finally encompassing to the entire protein when bound to a complete NTS. To further evaluate the inter-dependent coupling between protein domains, we computed the per-domain GC_{ij} scores (Cs), which accumulate (and normalize) the GC_{ij} for each protein domain with each other (details are reported in the SI).[93] As a result, the per-domain Cs matrices highlight the high degree of coupling between Nuc and REC2 upon DNA binding (Supplementary Figure A.13). AsCas12a, which has been simulated only in the DNA-bound state, also shows high GC_{ij} between REC1-2 and Nuc (Supplementary Figure A.11). The RNA-bound LbCas12a preserves a pattern of highly correlated motions (Supplementary Figure A.12). However, for these species, a direct comparison between

the RNA- and DNA-bound states is not possible due to the lack of structural information. Finally, it is interesting to note that only in the presence of complete nucleic acids (i.e., in the FnCas12a', Figure 2.4c), highly correlated motions are observed across the entire protein. This has also been established through MD simulations of the DNA-bound Cas9,[75,89,94] indicating a mechanism of interdomain allosteric communication.[95,96] This novel hypothesis in CRISPR-Cas12a arising from our data now warrants further computational investigations of protein allostery that are currently being pursued in our laboratory. This will likely also clarify the most important residues sustaining the functional motions and the communication mechanism.[97,98]

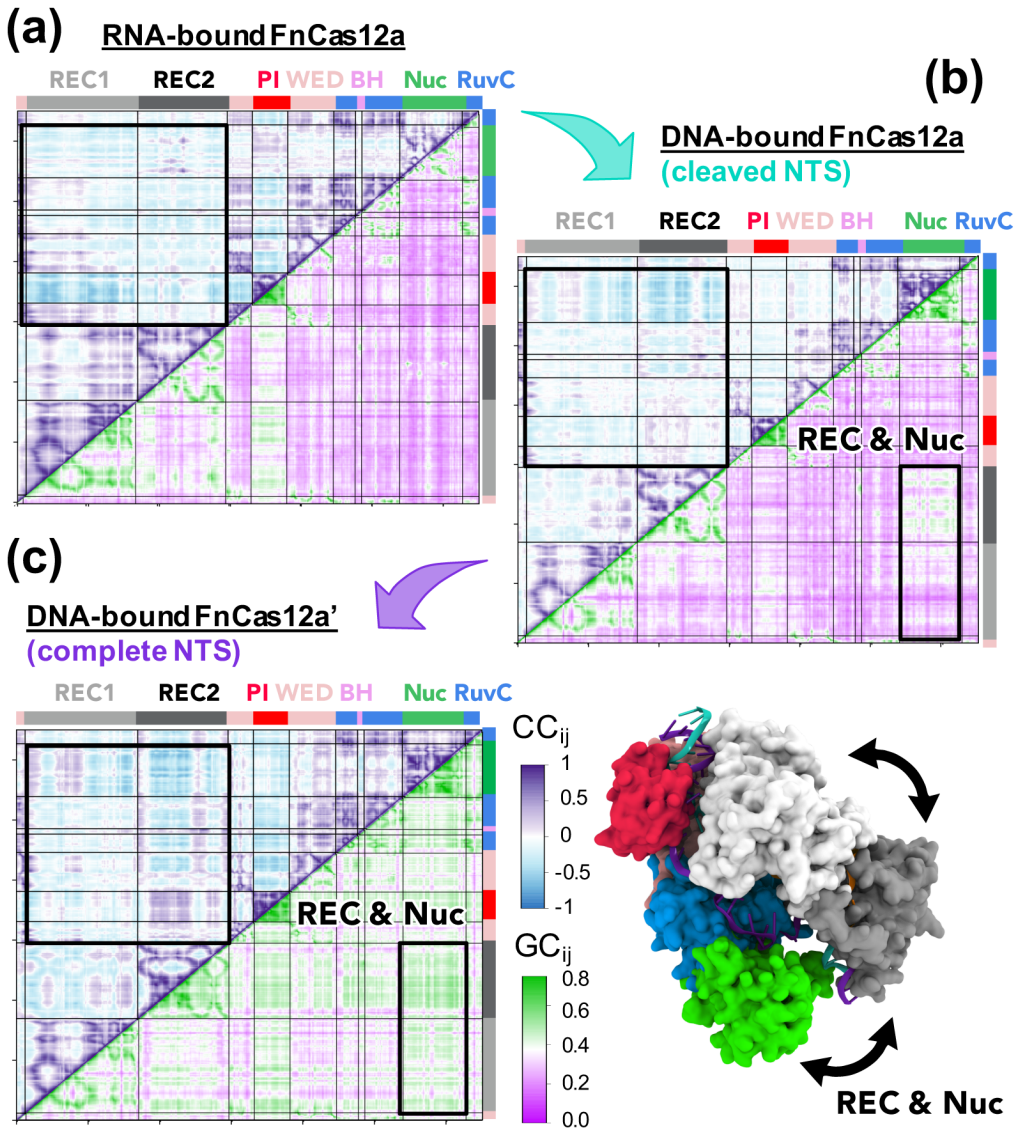


Figure 2.4: **Correlated motions of CRISPR-Cas12a.** Cross-Correlation (CC_{ij}, upper triangles) and Generalized Correlations (GC_{ij}, lower triangles) matrices, computed for FnCas12a in the RNA-bound state **(a)** and upon DNA binding, in the presence of a cleaved NTS **(b)** and a complete NTS **(c)**. Data are averaged over 4 simulation replicas of $\sim$1 μs each. The strength of the CC_{ij} is colored blue (for $CC_{ij} \geq 0$, lockstep motions) to violet (for $CC_{ij} \leq 0$, anticorrelated motions), while the GC_{ij} are green (correlated) to magenta (not correlated). Color scales are at the bottom. The protein sequence is also shown. Boxes are used to highlight anticorrelated CC_{ij} motions between the REC and NUC lobes, and highly coupled GC_{ij} between REC & Nuc. The latter are also shown on a cartoon of FnCas12a.

2.4 Discussion

2.4.1 Dynamic "Switch" upon DNA Binding

MD simulations of Cas12a indicate a change in the structural flexibility and in the conformational dynamics of the protein upon DNA binding. In the RNA-bound states, the PI region displays high flexibility (Figure 2.2) and large–amplitude motions directed toward the cleft that accommodates the PAM of the substrate DNA (Figure 2.3a). Upon DNA binding, the flexibility of the PI region is notably reduced, while the flexibility of REC2 and Nuc increases. Furthermore, the PI domain reduces the extent of its motions, while REC2 and Nuc display the largest amplitude motions. This suggests that the binding of DNA, which is initiated at the level of the PI domain in Cas12a, quenches the conformational flexibility of PI and induces the activation of peripheral large-scale motions at the level of REC2 and Nuc. This high flexibility and large–scale breathings of REC2 and Nuc upon DNA binding are also observed in the FnCas12a' complex, where REC2 and Nuc directly interact with a duplex region of the DNA substrate (Figure 2.1c). Taken together, analysis of the protein fluctuations (Figure 2.2) and large–amplitude motions (Figure 2.3) indicate a switch in the Cas12a dynamics upon DNA binding, which results in quenching of PI domain mobility and activating motions of the peripheral REC2 and Nuc domains. This switch in the Cas12a dynamics rationalizes some functional aspects and previous observations. Indeed, the high flexibility of the PI domain prior to DNA binding could favor its inward conformational change for accommodating the PAM.[40,72] Subsequent DNA binding which is initiated at the level of the PI region, results in curtailment of the PI domain dynamics. On the other hand,

REC2 and Nuc increase in flexibility, also when binding to a complete NTS (as in Fn-Cas12a', Figures 2.2 and 2.3). This is notable because the binding of dsDNA commonly induces the stabilization of the surrounding protein framework, which is indeed observed at the level of the PI domain. The high flexibility of REC2 and Nuc in the presence of complete nucleic acids also agrees with a very recent study revealing that the DNA in this region is intrinsically highly flexible.[58] Moreover, a high degree of flexibility of these regions upon DNA binding and during the formation of an RNA:DNA hybrid has also been observed through cryoEM and single-molecule experiments.[1,99] This increase in flexibility upon DNA binding is possibly a consequence of the need to cleave both DNA strands by a single active site located within the RuvC domain. This requires an exchange of the NTS and TS within the active site, which in turn necessitates conformational changes (and increased mobility) of the interacting protein domains, namely the REC lobe and Nuc.[1,48] Moreover, high conformational plasticity of REC and Nuc is also needed for the release of the product,[55] as well as for enabling the non-specific cleavage of single-stranded DNA (ssDNA) accessing the RuvC–Nuc interface.[16,59,68] This is a critical property that allows leveraging of Cas12a for the diagnosis of viral nucleic acids in the DETECTR technology.[17]

The switch in the conformational dynamics of Cas12a is also reflected by the analysis of the generalized correlations (GC_{ij}), which capture the overall dependencies of the protein motions. This analysis shows a change in the correlated motions of FnCas12a domains, revealing that correlations between REC1-2 and Nuc increase upon DNA binding (Figure 2.4, lower triangles). Notably, correlated motions of REC2 and Nuc are preserved across the DNA-bound Cas12a systems (Supplementary Figure A.8 -

A.12), suggesting a coupled dynamical function, where the motions of REC2 allow conformational changes in Nuc. This observation is particularly interesting in light of similar observations previously reported for the Cas9 enzyme. Indeed, in the DNA-bound form of Cas9, REC2 has been shown to play a critical role in regulating the conformational activation of the HNH domain for TS cleavage. Specifically, single-molecule experiments[79,81,83] and recent cryo-EM structures[80] have shown reciprocal conformational changes of REC2 and HNH domains to allow the latter to dock at the TS for cleavage. Accordingly, MD simulations have revealed that the HNH domain approaches the TS in concert with a transition of REC2,[75] while displaying correlated motions similar to what is observed for REC2 and Nuc in Cas12a. This suggests that despite being evolutionarily different than Cas9,[39] REC2 of Cas12a could play a similar regulatory function on the conformational activation of Nuc to allow the cleavage of the TS.

2.4.2 Conformational Activation for TS Processing

Cas12a uses the single RuvC domain to cleave both DNA strands, differing from Cas9, which cleaves the NTS and the TS using the RuvC and HNH nuclease domains, respectively. This has questioned the role and function of the additional Nuc domain, which was initially thought to cleave the DNA TS.[41] It is unclear, in fact, whether and how conformational changes of Nuc would activate DNA cleavages, or facilitate the exchange of the TS and NTS within the RuvC active site for cleavage.

Here, molecular simulations of Cas12a upon DNA binding offer interesting insights on the dynamic role of Nuc. We observed significant differences in the direction of the

essential motions of Cas12a in the presence/absence of the DNA NTS (Figure 2.3a). Indeed, in the presence of a complete NTS (i.e., in FnCas12a'), REC1-2 and Nuc display overall outward motions, leading to an expansion of the RuvC binding groove (Figure 2.5a). On the other hand, upon cleavage of the NTS (i.e., in AsCas12a and FnCas12a), REC1-2 and Nuc point inward (Figure 2.5b). This results in a contraction of the groove between the TS and the RuvC active site after NTS cleavage, which is also confirmed through volumetric analysis performed on the equilibrium trajectories (Supplementary Figure A.6). This observation is of particular interest. Considering that Cas12a cuts the NTS first,[1,2,42] this reinforces the outcomes of single-molecule FRET,[1] suggesting that Nuc and REC2 are highly dynamic and could allow the cleavage of the TS. Indeed, given the presence of a single catalytic site within the RuvC domain, after initial cleavage of the NTS, the TS should access the active site with conformational changes of the interacting protein domains. In light of this fact, our atomistic simulations indicate the tendency of Nuc to bend toward REC2, suggesting the narrowing of the groove between the TS and the RuvC active site. Concurrently, the motions of REC2 move in the opposite direction, showing the propensity to push the DNA TS toward the RuvC active site. REC1 also points downward, contributing the access of the TS to the active site. Hence, the REC lobe and Nuc would cooperate in the activation of the system toward TS cleavage.

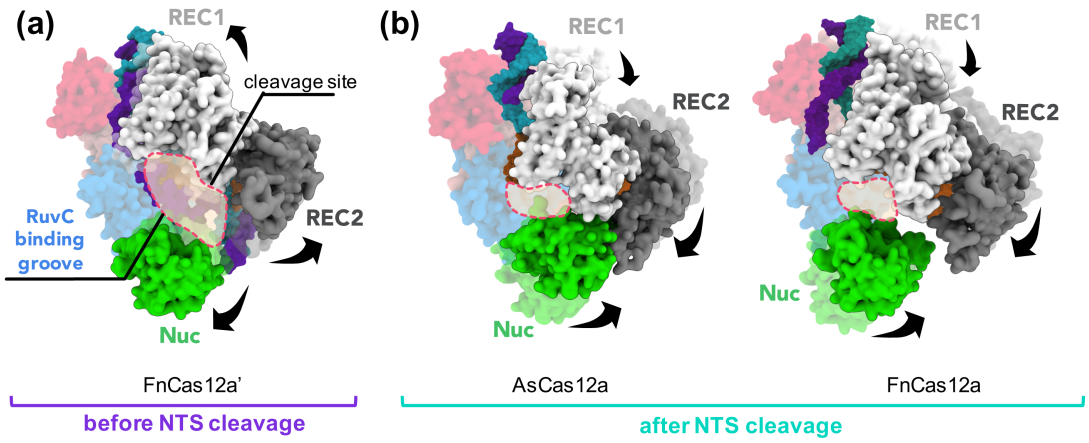


Figure 2.5: **Changes in the RuvC binding groove of CRISPR-Cas12a before and after NTS cleavage.** **(a)** Before NTS cleavage, an expansion in the RuvC binding groove is observed as mediated by outward motions of REC1-2 and Nuc. **(b)** After NTS cleavage, inward motions of REC1-2 and Nuc dynamics lead to a contraction of the RuvC binding groove. The outward/inward motions of the REC1-2 and Nuc domains are shown in transparent-to-mat colors. Arrows are also used to indicate the conformational transitions. The RuvC binding groove, which is located in between the REC and NUC lobes, is also highlighted.

Correlation analyses have indicated highly coupled motions between REC2 and Nuc (Figure 2.4). This suggests a coupled dynamical function, where the motions of REC2 are fundamental to allow the conformational changes of Nuc. This resembles the reciprocal dynamical role of REC2 and HNH in Cas9, where REC2 regulates conformational changes of the catalytic HNH domain to allow TS cleavage.[75,79–81] In light of these observations, the dynamics of REC2 in Cas12a could assist the conformational changes of Nuc and act as a "regulator" of its function. In support of this hypothesis, single-molecule studies of Cas12a have shown that the REC2 and Nuc domains show the largest conformational rearrangements.[1] Hence, the coupled dynamics observed here between REC2 and Nuc now calls for new biophysical experiments to assess the role of the REC lobe in the conformational activation of CRISPR-Cas12a. In this respect, we note that in the case of Cas9, biophysical studies have shown that conformational changes of the

REC lobe are also critical for the enzyme's specificity,[81,83] and mutations in the REC domain can reduce its off-target activity.[65,81,84] Moreover, a recent study has shown that point mutations in the REC lobe of AsCas12a can reduce off-target effects.[85] These findings and the results of the current study thus motivate further investigation on the functional role of REC and Nuc in CRISPR-Cas12a.

2.5 Conclusions

Here, all-atom MD simulations characterize the structural plasticity of CRISPR-Cas12a and the dynamic determinants underlying nucleic acid association. On a collective sampling of $\sim$20 μs, carried out over multiple states and across different species, we reveal that DNA binding induces a switch in the conformational dynamics of Cas12a, which results in quenching motions of the PAM interacting domain, and activating motions of the peripheral REC2 and Nuc domains. This switch in the Cas12a dynamics rationalizes crucial functional aspects. Indeed, the increased flexibility of REC2 and Nuc upon DNA binding could enable the conformational changes associated with DNA cleavage.[1] Considering that the core of the RuvC catalytic domain is highly rigid,[73,86] the observed flexibility of the adjacent Nuc domain and REC2 could contribute the exchange of the DNA strands within the Cas12a active site for sequential cleavage,[1] rapid release[55] and subsequent non-specific cleavage of ssDNAs.[59] This property is at the core of the DETECTR technology for rapid viral nucleic acid detection, especially that of SARS-CoV-2.[17] This suggests that the mutual dynamics of REC2 and Nuc

observed here could be critical for the non-specific binding of ssDNAs and thereby for the underlying mechanistic functioning of the DETECTR technology.

In-depth analysis of the dynamics upon DNA binding also offers mechanistic insights into the role of Nuc, whose function is incompletely understood. Accordingly, the joint dynamics of REC and Nuc shows the tendency to promote the conformational transition of the DNA TS toward the RuvC active site through opposite and concerted motions. REC2 and Nuc also display highly coupled dynamics, suggesting that REC2 could act as a regulator of the Nuc function, as previously observed for the HNH domain in Cas9.[75,79–81,83] In case of Cas9, the REC lobe play a critical role in the enzyme's activation and specificity.[65,81,84] Moreover, point mutations in REC of AsCas12a can reduce off-target effects,[85] which limit the applicability of the CRISPR technology. In light of these results, our outcomes motivate future investigations to characterize the functional role of REC and Nuc in CRISPR-Cas12a. It is also notable that our data suggest an unforeseen allosteric communication in CRISPR-Cas12a, which we have previously described in CRISPR-Cas9,[89,94,95] in agreement with experimental data.[44,81] This hypothesis grants now in-depth investigations, which we are currently pursuing building on our interests in the computational determination of protein allostery.

Overall, our work provides an atomic-level characterization of the CRISPR-Cas12a conformational dynamics, with insights into substrate DNA binding and cleavage. The mechanistic understandings arising from molecular simulations are of fundamental importance for further experimental studies aimed at a full characterization of the dynamic features of Cas12a. These outcomes can contribute to engineering efforts

aimed at improving the CRISPR-Cas12a technology toward more efficient and specific genome editing and viral detection.

2.6 Materials & Methods

2.6.1 Structural Models

Molecular simulations have been performed on five model systems of CRISPR-Cas12a, based on the available X-ray structures. The RNA-bound states have been based on two X-ray structures: *Lachnospiraceae bacterium* Cas12a (LbCas12a) solved at 2.38 Å resolution (5ID6.pdb)[43] and *Francisella novicida* Cas12a (FnCas12a) solved at 3.34 Å resolution (5NG6.pdb).[2] The DNA-bound states have been based on three X-ray structures. We considered the structures of *Acidaminococcus sp.* Cas12a (AsCas12a) and FnCas12a, in which the NTS is partially cleaved (i.e., 5B43.pdb at 2.38 Å resolution[41] and 5NFV.pdb at 2.50 Å resolution[2] respectively). The third DNA-bound state has been based on a recent structure of the FnCas12a, including a longer NTS that binds within the RuvC cleft and reconciles with the TS (6I1K.pdb at 2.65 Å resolution).[16] These systems have been embedded in explicit waters and Na+ ions were added to neutralize the total charge, leading to orthorhombic periodic cells comprising on average a total number of 210,000 atoms, for each system. Full details are reported in the Supporting Information (SI).

2.6.2 Molecular Dynamics (MD) Simulations

MD simulations have been performed employing Amber ff12SB force field, which includes the ff99bsc0[100] corrections for DNA and the ff99bsc0+χOL3[101,102] corrections for RNA. Allnér force field[103] has been employed for Mg^{2+} ions and TIP3P model[104] for waters. These force field parameters and the simulation protocol have also been employed in our recent studies of CRISPR-Cas9,[94,105,106] corroborated by NMR experiments and quantum mechanical calculations, enabling a fair comparison. An integration time step of 2 fs has been employed. All bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm. Temperature control (300 K) has been performed via Langevin dynamics,[107] with a collision frequency γ = 1/ps. Pressure control was accomplished by coupling the system to a Berendsen barostat,[108] at a reference pressure of 1 atm and with a relaxation time of 2 ps. The systems have been subjected to energy minimization to relax water molecules and counter ions, keeping the protein, RNA, DNA and Mg^{2+} ions fixed with harmonic position restraints of 300 kcal/mol $\cdot$ $\mathring{A}^2$. Then, the systems have been heated up from 0 to 100 K in the canonical ensemble (NVT), by running two simulations of 5 ps each, imposing position restraints of 100 kcal/mol $\cdot$ $\mathring{A}^2$ on the above-mentioned elements of each system. The temperature was further increased up to 200 K in $\sim$100 ps of MD runs in the isothermal-isobaric ensemble (NPT), reducing the restraint to 25 kcal/mol $\mathring{A}^2$. Subsequently, all restraints were released and the temperature of the systems was raised up to 300 K in a single NPT simulation of 500 ps. After $\sim$1.1 ns of equilibration, $\sim$10 ns of NPT runs were carried out allowing the density of the system to stabilize around 1.01 g/cm^{-3}. Finally, $\sim$1 μs of MD simulations has been carried out in an NVT ensemble for each system, which has

also been simulated in 4 replicates. Independent MD simulations replicas have been obtained starting from different configurations and velocities, initialized accordingly to the Maxwell-Boltzmann distribution at physiological temperature. This approach enabled us to obtain solid statistics for the analysis in our purposes. Considering five simulation systems, we collected a total of $\sim$20 μs of aggregate sampling (i.e., 5 systems * 4 replicas * $\sim$1 μs = $\sim$20 μs). Molecular simulations have been performed using the GPU version of AMBER 18.[109] The analysis of the results has been performed on each simulated MD replica, and on the ensemble obtained averaging independent ns–to–μs trajectories (details are in the SI). This enabled assessing the reproducibility of our results across independent simulations, and also to provide a solid statistical ensemble.

2.6.3 Principal Component Analysis (PCA)

PCA is a statistical method that can report large-scale collective motions occurring in biological macromolecules undergoing MD simulations. Through this statistical technique, it is possible to reduce the large number of degrees of freedom to an essential subspace set, which captures large-amplitude motions of the system. In PCA, the covariance matrix of the protein Cα atoms is calculated and diagonalized to obtain a new set of coordinates (eigenvectors) to describe the system motions. Each eigenvector – also called Principal Component (PC) – is associated with an eigenvalue corresponding to the mean square fluctuation contained in the system's trajectory projected along that eigenvector. By sorting the eigenvectors according to their eigenvalues, the first PC (i.e., PC1) corresponds to the system's largest amplitude motion, and the dynamics of the system along PC1 is usually referred to as "essential dynamics".[87]

In this work, the principal motions of the protein were captured starting from the mass-weighted covariance matrix of the Cα atoms. In detail, PCA has been performed considering the FnCas12a systems, whereby the collected ensembles (i.e., arising from the compared RNA-bound FnCas12a, DNA-bound FnCas12a and FnCas12a' systems; Figure 2.3) were combined and subjected to RMS-fit to the same reference configuration, removing the rotational and translational motions. This was performed to ensure a consistent eigenbasis and motions of the PCs on all compared systems and to construct the covariance matrices from the atoms' positions. We also performed two independent PCA on the LbCas12a and AsCas12a systems, providing insights into their essential motions. Each element in the covariance matrix is the covariance between atoms i and j, defining the i, j position of the matrix. The covariance C_{ij} is defined as:

$$C_{ij} = \langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle \tag{2.1}$$

where $\vec{r}_i$ and $\vec{r}_i$ are the position vectors of atoms i and j, and the brackets denote an average over the sampled time period. The two terms in Equation 2.1 represent the displacement vectors for atoms i and j. The covariance matrix was then diagonalized, leading to a complete set of orthogonal collective eigenvectors, each associated to a corresponding eigenvalue. The eigenvalues denote how much each eigenvector is representative of the system dynamics, thus giving a measure of the contribution of each eigenvector to the total variance. Indeed, the eigenvectors with the largest eigenvalues correspond to the most relevant motions. By projecting the displacements vectors of each atom along the trajectory onto the eigenvectors (i.e., by taking the dot product between the two vectors

at each frame), the PCs were then obtained. The cumulative variance accounted by all the PCs was calculated for all systems, revealing that the first PCs account for the major contribution (Supplementary Figures A.5-A.6). Full details on the application of this statistical technique on MD simulations of CRISPR-Cas12a are in the SI.

2.6.4 Cross-Correlation analysis

Correlation analysis has been performed in order to identify the dynamical coupling of the motions between Cα atoms (i and j) in the simulated systems. "Pearson-like" cross-correlation (CC_{ij}) analysis provides a measure of the collinear correlations between the atoms i and j. The $CCij$ matrix can be computed as a normalization of the covariance matrix:

$$CC_{ij} = \frac{\langle(\vec{r}_i - \langle\vec{r}_i\rangle)(\vec{r}_j - \langle\vec{r}_j\rangle)\rangle}{[(\langle\vec{r}_i{}^2\rangle - \langle\vec{r}_i\rangle^2)(\langle\vec{r}_j{}^2\rangle - \langle\vec{r}_j\rangle^2)]} \tag{2.2}$$

where $\vec{r}_i$ and $\vec{r}_j$ are the position vectors of atoms i and j, considered over the sampled time period (denoted using brackets). Positive values of the CC_{ij} coefficients indicate lockstep motions between atoms i and j; while negative CC_{ij} values are indicative of anticorrelated motions. CC_{ij} values equal to zero evince that the atoms' displacements are independent from each other. The magnitude of CC_{ij} coefficients (i.e., ranging from 0 – 1 for lockstep motions; and from –1 – 0 for anticorrelated motions) indicates strength of the correlation. As noted above, the CC_{ij} neglect the non-linear contributions between atoms i and j; and do not capture correlated motions occurring out of phase with each other. To capture more broadly the dependency of the atomic motions, we also employed a generalized correlation method described below.

2.6.5 Generalized Correlations analysis based on Mutual Information

This approach relies on information theory and uses the mutual information (MI) measure to obtain the generalized correlation (GC_{ij}) coefficients.[92] In information theory, two variables, such as the such as the $\vec{r}_i$ and $\vec{r}_j$ position vectors, can be considered correlated when their joint probability distribution, $p(\vec{r}_i,\vec{r}_j)$, is smaller than the product of their marginal distributions, $p(\vec{r}_i)\cdot(\vec{r}_i)$. The MI is a measure of the degree of correlation between $\vec{r}_i$ and $\vec{r}_i$ defined as function of $p(\vec{r}_i,\vec{r}_j)$ and $p(\vec{r}_i)\cdot(\vec{r}_i)$ accordingly to:

$$MI[\vec{r}_i,\vec{r}_j] = \int\int p(\vec{r}_i,\vec{r}_j) ln \frac{p(\vec{r}_i,\vec{r}_j)}{p(\vec{r}_i)\cdot(\vec{r}_i)} d\vec{r}_i d\vec{r}_j \tag{2.3}$$

Notably, MI is closely related to the definition of the Shannon entropy and can be computed as:

$$MI[\vec{r}_i,\vec{r}_j] = H[\vec{r}_i] + H[\vec{r}_j] - H[\vec{r}_i,\vec{r}_j] \tag{2.4}$$

where $H[\vec{r}_i]$ and $H[\vec{r}_j]$ are the marginal Shannon entropies, and $H[\vec{r}_i,\vec{r}_j]$ is the joint entropy, providing a link between motions' correlations and information content. Based on this definition, and considering that MI varies from 0 to $+\infty$, the normalized GC_{ij} coefficients, ranging from 0 (independent variables) to 1 (fully correlated variables), can be defined as:

$$GC_{ij}[\vec{r}_i,\vec{r}_j] = \{1 - e^{-2MI[\vec{r}_i,\vec{r}_j]/d}\}^{-1/2} \tag{2.5}$$

where d is the dimensionality of $\vec{r}_i$ and $\vec{r}_i$. This approach has been originally introduced by Lange and Grubmüller,[92] who developed a computationally-efficient algorithm. Overall, a GC_{ij} analysis is powerful in capturing non-linear coupled motions in biomolecular systems. However, the GC_{ij} coefficients do not distinguish positive *vs.* negative motions, giving a normalized measure of how much information on one atom's position is provided by that of another atom. Hence, when employed together, the CC_{ij} and GC_{ij} schemes can provide a more comprehensive understanding of the interdependent dynamics of proteins, with information on whether protein regions move in lockstep or through opposite motions (through CC_{ij}), and providing also a more general information on the atoms' interdependence (through GC_{ij}). In the present work, the GC_{ij} coefficients have been computed using the positions vectors of Cα atoms along the simulated trajectories. For each model system, all correlation analyses have been performed over independent MD trajectories, and have also been averaged over the aggregate sampling arising from 4 ns–to–μs MD replicas (details are in the SI). To further spotlight relevant correlations among spatially distant domains, the averaged GC_{ij} matrices of the FnCas12a system have also been further processed to compute per-domain GC_{ij} scores (Cs). This measure accumulates and normalizes the GC_{ij} coefficients over each protein domain, resulting in per-domain GC_{ij} matrices that help in identifying the most relevant coupled motions in large biomolecular systems (details are in the SI).[73,93]

2.7 Acknowledgments

We thank Lukasz Nierzwicki for his thoughtful comments. This material is based upon work supported by the National Science Foundation under Grant No. CHE-1905374. This work was partially funded by NIH Grant R01 EY027440. This work used resources services, and support provided via the COVID-19 HPC Consortium (https://covid19-hpc-consortium.org/), which is a unique private-public effort to bring together government, industry, and academic leaders who are volunteering free compute time and resources in support of COVID-19 research. Computer time for MD simulations has also been awarded by XSEDE via the grant TG- MCB160059. We also acknowledge Microsoft Azure for granting us cloud computing resources for computations and analysis.

Chapter III

An Alpha-helical Lid Guides the Target DNA toward Catalysis in CRISPR-Cas12a

Reprinted with permission from Saha A. *et al.*

3.1 Abstract

CRISPR-Cas12a is a powerful RNA-guided genome-editing system that generates double-strand DNA breaks using its single RuvC nuclease domain by a sequential mechanism in which initial cleavage of the non-target strand is followed by target strand cleavage. How the spatially distant DNA target strand traverses toward the RuvC catalytic core is presently not understood. Here, continuous tens of microsecond-long molecular dynamics and free-energy simulations reveal that an α-helical lid, located within the RuvC

domain, plays a pivotal role in the traversal of the DNA target strand by anchoring the crRNA:target strand duplex and guiding the target strand toward the RuvC core, as also corroborated by DNA cleavage experiments. In this mechanism, the REC2 domain pushes the crRNA:target strand duplex toward the core of the enzyme, while the Nuc domain aids the bending and accommodation of the target strand within the RuvC core by bending inward. Understanding of this critical process underlying Cas12a activity will enrich fundamental knowledge and facilitate further engineering strategies for genome editing.

3.2 Introduction

Bacteria and other prokaryotes deploy an arsenal of CRISPR-Cas systems (clustered regularly interspaced short palindromic repeats, CRISPR-associated proteins) as a part of their adaptive immune response against foreign nucleic acid invasion.[11] The CRISPR-associated proteins Cas9[8] and Cas12a,[15] being RNA-programmable endonucleases, launched unparalleled genome-editing applications across different fields of biotechnology, agriculture, and medicine.[14,110] The Cas12a nuclease came to the forefront with innovative applications such as extraordinarily rapid nucleic acid detection,[59] also harnessed for the detection of the SARS-CoV-2.[17,111] At the molecular level, Cas12a cleaves double-stranded target DNA using a single catalytic domain.[1,2,16,40,43,72] Hence, conformational changes of the protein are critical for the accommodation and cleavage of DNA, but the mechanism of such a conformational transition is highly elusive.

Structures of Cas12a from *Francisella novicida* (FnCas12a) revealed a bilobed architecture, comprising of a recognition (REC) and a nuclease (NUC) lobe, flanked by the wedge (WED) domain (Figure 3.1).[1,2,16,40] Two alpha-helical domains (REC1 and REC2) constitute the REC lobe and mediate nucleic acid binding, while the NUC lobe comprises the RuvC catalytic domain and an adjacent Nuc domain. During its biophysical function, Cas12a uses CRISPR RNA (crRNA) for highly specific molecular recognition[48,55,65] and cleavage of complementary DNA sequences. Upon recognition and binding of a short protospacer-adjacent motif (PAM) in the target DNA by the PAM-interacting (PI) domain of Cas12a, the crRNA binds one DNA strand (i.e., the target strand, TS) and forms a crRNA:TS duplex, while the other non-target strand (NTS) gets displaced. Cas12a then confers double-stranded DNA breaks by using only the RuvC domain as its molecular scissor. This is strikingly different from Cas9, where two catalytic domains – HNH and RuvC – perform cleavage of the TS and NTS, respectively.[13] This difference alludes to a conundrum of how a single nuclease domain could enable cleavage of both DNA strands in Cas12a. Indeed, while the NTS accommodates within the RuvC catalytic cleft, the TS stays spatially distant from the catalytic pocket, close to the REC lobe (Figure 3.1). Experimental observations have inferred that the TS gets cleaved post the cleavage of the NTS.[1,2,42] Together, this indicates a conformational change after the NTS cleavage, allowing the accommodation of the TS within the RuvC catalytic pocket. Nevertheless, the molecular details of these intriguing biophysical processes have remained incompletely understood.

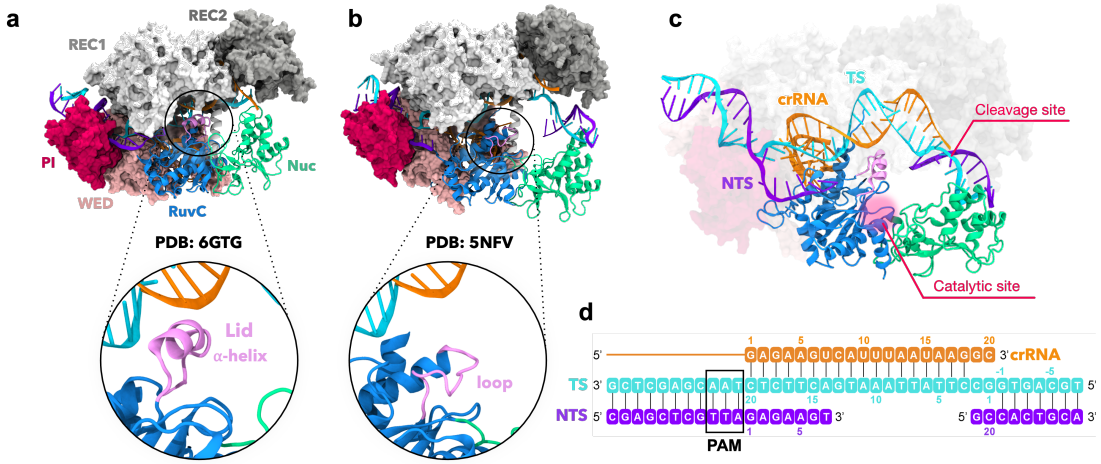


Figure 3.1: **Structures of CRISPR-Cas12a after DNA non-target strand cleavage.** The Cas12a protein is shown in molecular surface, highlighting the RuvC (blue) and Nuc (green) domains as ribbons. The crRNA (orange), the DNA target strand (TS, cyan) and the non-target strand (NTS, violet) are also shown as ribbons. **(a)** The cryo-EM structure (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation))[1] contains a short TS without the scissile phosphate. This structure displays an α-helical "lid", opening the RuvC domain (close-up view). **(b)** The X-ray structure (PDB: 5NFV [https://doi.org/10.2210/pdb6GTG/pdb] (FnCas12a bound to R-loop))[2] reports a longer TS, including the scissile phosphate and reconciling with the NTS. Here, the lid is unstructured (close-up view) and reconstructed through homology modelling. **(c)** A more complete CRISPR-Cas12a modelled from the cryo-EM **(a)** and X-ray **(b)** structures, containing the α-helical lid and a longer DNA TS including the cleavage site. **(d)** Schematic representation of the nucleic acids.

Alongside biochemical and structural studies,[1,2,51] single-molecule experiments[1,42,47,52,53,55,112] have substantially contributed toward delineating the biophysical functioning of the Cas12a enzyme. The RuvC domain is structurally stable in Cas12a, while the REC2 and Nuc domains display increased flexibility and concerted motions upon DNA binding, which could favour the conformational transition of the TS toward the catalytic site. Such cooperative movements of REC2 and Nuc have been demonstrated by cryo-EM and single-molecule FRET experiments,[1] and shown through molecular simulations.[113] However, the protein dynamics and the conformational transitions orchestrating the traversal of the TS from the REC lobe to the RuvC catalytic site

are yet to be identified. This is a critical question, understanding of which can lay the ground for further engineering toward improved genome editing.

Here, extensive Molecular Dynamics (MD) and free-energy simulations of Cas12a are used to unravel the process of DNA TS traversal and its repositioning within the RuvC catalytic core from a total simulation time of $\sim$170 μs. Using Anton2, one of the fastest supercomputers available for MD simulations,[114] we performed continuous tens of μs-long runs, identifying an α-helical "lid" as a key structural element, and the REC2 and Nuc domains as the major players orchestrating the conformational transition. Next, free-energy methods captured the elaborate mechanism by which the α-helical lid guides the DNA TS toward the catalytic core, as also supported by DNA cleavage experiments. Taken together, our findings shed light on a cardinal step in the functioning of Cas12a, which has previously remained highly elusive.

3.3 Results

3.3.1 An α-helical Lid Anchors the crRNA:TS

Our investigations considered two structures obtained through cryo-EM and crystallographic techniques upon NTS cleavage with different lengths of DNA TS (Supplementary Table B.1). The cryo-EM structure (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation); EMD-0065,[1] Figure 3.1a) contains a short TS without the scissile phosphate. This structure has an open RuvC domain, displaying an α-helical "lid" (residues L1008–K1021) that was held accountable for opening the RuvC domain for DNA accommodation.[1] The X-ray structure

(PDB: 5NFV [https://doi.org/10.2210/pdb6GTG/pdb] (FnCas12a bound to R-loop),[2] Figure 3.1b) displays a full-length TS with the scissile phosphate, while the lid is structurally disordered.

Our $\sim$10 μs long simulations of the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) cryo-EM structure suggest that the α-helical lid plays a crucial role in the major conformational changes of the system. Indeed, the α-helical lid draws the TS by establishing extensive interactions between its Lys/Arg residues and the crRNA:TS duplex backbone at PAM-distal sites (Supplementary Figure B.1). It is notable, however, that in this cryo-EM structure, the region of the TS downstream of the crRNA:TS duplex is absent and consequently lacks the scissile phosphate.[1] Hence, we modeled the downstream region of the DNA in the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) structure based on the 5NFV [https://doi.org/10.2210/pdb6GTG/pdb][2] (FnCas12a bound to R-loop) X-ray structure, obtaining a more complete system of Cas12a after NTS cleavage (Figure 3.1c; more details in the methods section). Two $\sim$10 μs long replicates of this system consistently show that positively charged residues of the α-helical lid interact with the crRNA:TS duplex at PAM-distal sites after $\sim$2 μs of MD (Figure 3.2a-b). These include the stable interaction of K1013 with the TS at position 2, and R1016 with the crRNA at position 12. A bending of the TS at the end of the crRNA:TS duplex is also observed that becomes more prominent after $\sim$5.2 μs in both simulation replicates, leading to an $\sim$40° bending with respect to the initial state (Figure 3.2c). This is an important observation, which is in line with the experimental evidence of a distortion of the TS downstream of the crRNA:TS duplex that was deemed necessary for it to reach the

RuvC cleavage site.[58] Remarkably, after $\sim$5.8 μs of MD, the region of the TS including the scissile phosphate also moves toward the RuvC active site in both MD runs (Figure 3.2d), further supporting the experimental indications. It is notable that biochemical and single-molecule experiments have shown that the cleavage site on the TS can mainly be located between base positions -2 and -4.[53,58] Hence, we measured the distance between the scissile phosphate connecting the T-2 and G-3 nucleobases and the center of mass of the RuvC catalytic core (Figure 3.2d, Supplementary Figure B.2d,e). The RuvC catalytic core is composed of the D917, E1006, and D1255 residues coordinating two Mg^{2+} ions, displaying a conserved two-metal ion architecture similar to Cas9.[78,115] During our $\sim$10 μs long replicates, a substantial decrease in the computed distance indicates that the scissile phosphate approaches the RuvC core.

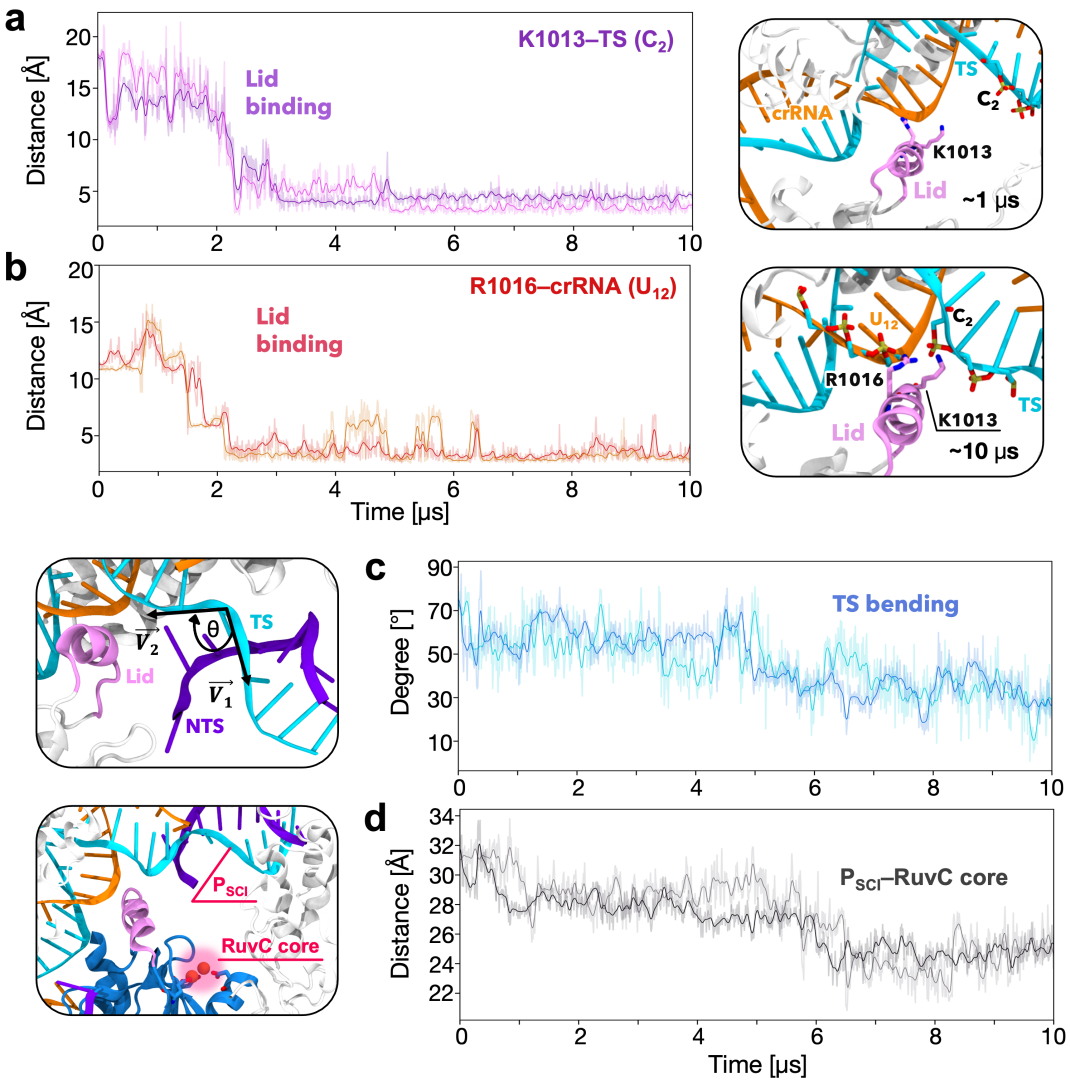

a
Distance [Å]
K1013–TS (C2)
Lid
binding
TS
C2
crRNA
K1013
Lid
~1 μs
b
Distance [Å]
R1016–crRNA (U12)
Lid
binding
U12
C2
R1016
TS
K1013
Lid
~10 μs
Time [μs]
V2
θ
TS
Lid
V1
NTS
c
Degree [°]
TS bending
PSCI
RuvC core
d
Distance [Å]
PSCI–RuvC core
Time [μs]

Figure 3.2: **The α-helical lid anchors the crRNA:TS duplex.** **(a)** Interactions established by the positively charged residues of the α-helical lid with the DNA target strand (K1013:NZ – C_2:OP1), and **(b)** the crRNA (R1016:NH1 – U_{12}:OP1), evolving over $\sim$10 μs of molecular dynamics simulations in two replicates. Interactions are measured as oxygen-nitrogen distance and are considered formed under 4 Å distance.[116] Representative snapshots on the right show the interactions between the α-helical lid (mauve) and the crRNA:TS duplex (crRNA: orange, TS: cyan) at different simulation times. **(c)** Bending of the DNA TS, computed as angle θ between the vectors ($\vec{V}_1$) and ($\vec{V}_2$) passing through G_{-1}, G_1, C_2, C_3 and T_4 (left panel, details in the Supplementary Methods). After $\sim$5 μs of MD, the region of the TS at the end of the duplex bends, leading to an $\sim$40° bending with respect to the initial state. **(d)** Distance between the scissile phosphate (P_{SCI}, between the T_{-2} and G_{-3} nucleobases) and the center of mass of the RuvC catalytic core (left panel, composed of D917, E1006 and D1255 residues coordinating with two Mg^{2+} ions). P_{SCI} moves toward the RuvC core after $\sim$5.8 μs of MD. Data are reported for two simulation replicates of $\sim$10 μs each. The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation.

Taken together, the two $\sim$10 μs long simulations of Cas12a, including a longer TS and the scissile phosphate, confirm that the α-helical lid anchors the crRNA:TS duplex at PAM-distal sites. Upon these interactions, the TS bends and the scissile phosphate moves toward the RuvC core. The observed interactions thereby suggest that the lid could have a critical role in the traversal of the DNA TS for catalysis.

The 5NFV [https://doi.org/10.2210/pdb6GTG/pdb][2] (FnCas12a bound to R-loop) X-ray structure captured the protein complex after NTS cleavage, without a structured lid.[2] Homology modeling reconstructed this disordered lid into an unstructured loop (Figure 3.1b, see the Methods section). Molecular simulations ($\sim$10 μs in two replicates) of this system show very high flexibility of the unstructured loop, especially in comparison with the α-helical lid (Supplementary Figure B.3). Interestingly, this unstructured lid forms electrostatic interactions with the crRNA but fails to establish persistent interactions with the TS (Supplementary Figure B.4a-b). These interactions between the lid and the crRNA:TS duplex are slightly different (adjacent bases) compared to the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) cryo-EM structure, mostly likely owing to the differences in the lengths of the TS. Thus, though the TS comes closer to the RuvC catalytic site (Supplementary Figure B.4d), most likely owing to the inward movement of REC2 and the arching of the crRNA-TS duplex toward RuvC (Supplementary Figure B.5), yet the TS fails to bend (Supplementary Figure B.4c). Considering that TS bending was deemed crucial for its accommodation at the RuvC active site,[58] an unstructured lid is thus functionally less effective for the accommodation of the TS.

3.3.2 REC2 and Nuc Facilitate the Traversal of the TS

Our long-timescale simulations of CRISPR-Cas12a including a complete TS (Figure 3.1c) also reveal noteworthy conformational changes of the protein and the crRNA:TS duplex, suggesting a cooperative dynamical mechanism fostering the traversal of the TS. To monitor these conformational changes, we analyzed the bending angles between the vector passing through the regions of interest and its perpendicular out of the plane, with respect to the first frame of reference (details in the Supplementary Methods).

We observe that the long α-helices of REC2 binding the crRNA:TS minor groove (residues 339-383, 474-543, 557-590; hereafter referred as REC2a) exhibit restricted motions. On the other hand, the short alpha helices connected by turns (residues 384-473, 543-557; referred to as REC2b) display a marked inward movement along the two $\sim$10 μs long simulations. This is shown by a decrement in the bending angle of REC2b (Figure 3.3a, gray), which is noteworthy at $\sim$3 μs and after $\sim$6 μs, with REC2b becoming more bent in one simulation replica than the other after $\sim$6 μs (as detailed in regular time-series plots in Supplementary Figure B.6a). This conformational change follows a similar trend of the arching of the crRNA:TS duplex (Figure 3.3a, orange and Supplementary Figure B.6b). Indeed, the terminal major groove of the duplex, adjacent to REC2b, arches toward the RuvC domain, as shown by the reduction of its vector angle in both the simulation replicates (Supplementary Figure B.6b). The duplex arching is remarkable $\sim$6 μs of MD onward, alongside REC2b bending. The cooperativity between the REC2b bending and hybrid arching is also shown by the positive correlation between the two variables (Supplementary Figure B.6f). Additionally, REC2b maintains extensive contacts with the crRNA:TS duplex throughout the simulations (Supplementary

Figure B.7a). This indicates that the inward motion of REC2b helps the arching of the crRNA:TS duplex. Both conformational changes are observed conspicuously upon the α-helical lid binding to the crRNA:TS duplex, occurring at ∼2 μs in both simulation replicates (Figure 3.2). These observations suggest the cooperativity between REC2 and the α-helical lid. Indeed, the α-helical lid anchors the crRNA:TS duplex, which is followed by the bending of REC2, facilitating conformational changes of the duplex toward the RuvC catalytic pocket.

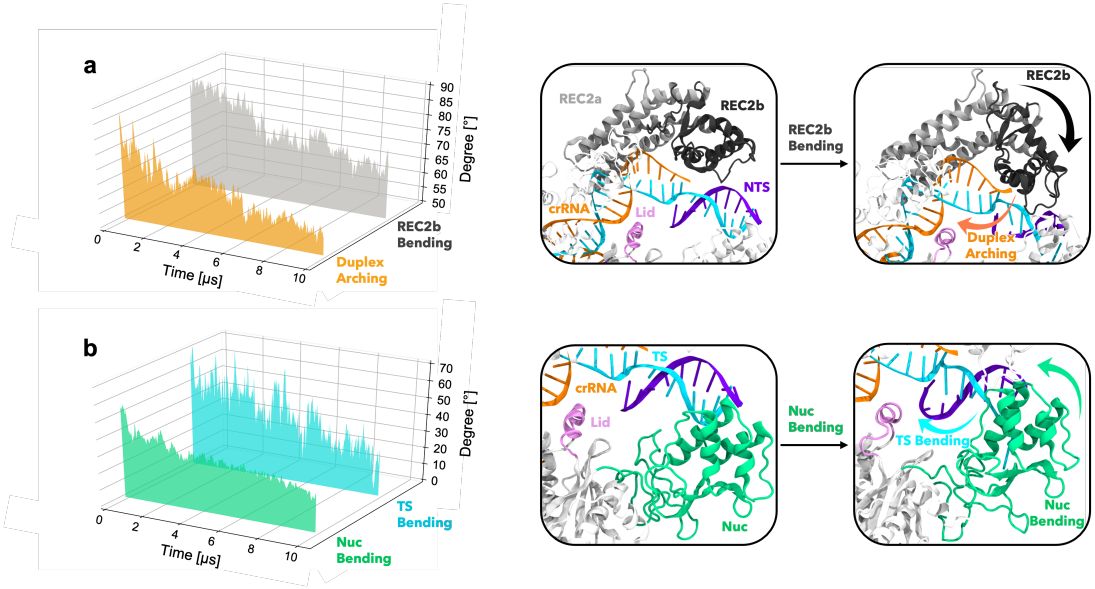


Figure 3.3: **Bending of major domains in CRISPR-Cas12a.** **(a)** Bending of the REC2b region (residues 384-473, 543-557, gray) and arching of the terminal major groove of the crRNA:TS duplex (base pairs A13:G19–T8:C2) along ∼10 μs of MD simulations of CRISPR-Cas12a including a complete TS. The REC2b and crRNA:TS duplex conformational changes are shown on the right. **(b)** Bending of the DNA TS (cyan, Figure 3.2d) in concert with the bending of the Nuc domain (green) evolving over the multi-μs simulation. The panels on the right show the conformational changes. The curved arrows indicate the conformational changes in the domains. Details are in the Supplementary Methods and Supplementary Table B.2. Data are shown here for one simulation replica, and for two replicates in Supplementary Figure B.6, in which regular time-series plots are reported.

Our simulations also show that the Nuc domain bends toward the core of the enzyme, as evidenced by the decrement of the Nuc bending angle (Figure 3.3b, green and Supplementary Figure B.6c reporting regular time-series plots for two simulation replicas). During this conformational transition, Nuc maintains a noteworthy number of contacts with the 5'-region of the TS (Supplementary Figure B.7b). The DNA TS at the end of the duplex also bends in both the simulation replicates (Figure 3.3b, cyan and Supplementary Figure B.6d). This similar trend of the Nuc and TS bending angles, supported by a positive correlation between the two variables (Supplementary Figure B.6g), along with the persistent interactions between them suggest that the bending of Nuc aids in the bending of the DNA TS.

Cryo-EM and single-molecule experiments reported high flexibility and coupled dynamics of REC2 and Nuc, suggesting these conformational changes to be crucial for DNA cleavage.[1,47,52,53,55,112] The continuous multi-μs long simulations reported here demonstrate that REC2 and Nuc bend toward the core of the enzyme (Figure 3.3 and Supplementary Figure B.6e), while establishing multiple contacts between them (Supplementary Figure B.7c). These large-scale coupled motions exhibited by the major domains, along with the lid anchoring the TS, offer a compact conformation to the whole protein, facilitating the traversal of the TS toward the catalytic pocket.

3.3.3 Traversal of the DNA TS and Accommodation in the RuvC Catalytic Pocket

The region of the TS including the scissile phosphate is significantly far from the catalytic core ($\sim$30 Å, Figure 3.1), which opens an overarching question regarding

its traversal and accommodation within RuvC. To investigate this phenomenon, we performed free-energy simulations, sampling the dynamics of the complex from the complete Cas12a structure (Figure 3.1c) to a catalytically competent state (Supplementary Figure B.8). The latter was built based on related X-ray structures of Cas12a orthologues holding the DNA TS within the RuvC active site (see Supplementary Methods). This catalytically competent state was subjected to $\sim$10 μs long MD simulations to ensure extensive relaxation, revealing that no further major conformational change happens in the long timescale, especially in the dynamics of REC2 and Nuc (Supplementary Figure B.9a). This is indeed an important point, considering that the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) and 5NFV [https://doi.org/10.2210/pdb6GTG/pdb][2] (FnCas12a bound to R-loop) structures underwent major conformational changes over $\sim$6 – 7 μs of continuous MD (Figures 3.2-3.3). Importantly, the simulations also show that the RuvC active site with the DNA TS is remarkably stable (Supplementary Figure B.9b-d).

We performed Umbrella Sampling (US) simulations,[117] studying the traversal of the DNA TS along two reaction coordinates (RCs). Here, the complete Cas12a structure (Figure 3.1c) with the full-length TS close to the REC2 domain was the initial state, while the catalytically competent Cas12a structure obtained after $\sim$10 μs-long MD simulation (Supplementary Figure B.8c) was the final state. Structural,[1,2,16,40,43,72] single-molecule[1,42,47,52,53,55,112] and computational[113] studies suggested that conformational changes in the Cas12a protein are needed to conduct the traversal of the TS. Hence, the difference in root-mean-square deviation (RMSD) of all the Cα atoms between the initial and the final states appealed as an appropriate RC (RC1) to capture the

slowest dynamics involved in the conformational transition. Additionally, the second RC (RC2) was selected as the distance between the center of mass of the RuvC catalytic core and that of the TS region including the scissile phosphate (i.e., the DNA nucleobases at positions -2 to -4, as indicated by biochemical and single-molecule data).[53,58] The conformational transition was sampled over $\sim$50 μs. Upon removing the biases incurred by the restraints, the free energy landscape (i.e., the Potential of Mean Force, PMF) was described along its minimum free energy pathway (see the Methods section, Supplementary Figures B.10-B.12).

We observe that the free energy profile (Figure 3.4a, blue line) flows freely through a plateau (i.e., state A), barrierless toward the first minimum (state B) as the traversal of the DNA TS is fostered by the interactions of the α-helical lid with the crRNA:TS duplex. Indeed, the α-helical lid anchors the duplex with its positively charged residues (K1013, R1014, R1016, K1018), while interacting with the nucleobases using N1009 (Supplementary Figure B.13a,c-f). A phenylalanine residue (F1012) is also observed to establish stacking interactions with the nucleobases of the TS (with G_{-1} and T_{-2}; Supplementary Figure B.13b). These critical interactions with the lid residues guide the descent of the TS. The α-helical lid shows high dynamicity as it traverses through state A (Supplementary Figure B.14a) and restores α helicity again along the minimum

free energy path (Supplementary Figure B.14a-c) as the profile slides down to state B and establishes extensive network of interactions with the crRNA:TS duplex (Supplementary Figure B.13). Then, a $\sim$3 kcal/mol barrier is observed leading to the second minimum (state C) corresponding to the DNA TS repositioning within the catalytic core, and the closure of the Nuc domain toward the core of the enzyme (Figure 3.4a and Supplementary Figure B.15c-d), constituting the slowest step in the process. These observations suggest that Nuc is critical in aiding the accommodation of the DNA TS within the RuvC catalytic core. Accordingly, the K1206-R1218 α-helix/β-sheet region of Nuc maintains persistent interaction with the 5'-tail of the DNA TS (Supplementary Figure B.15e). Hence, Nuc could hold the TS within the RuvC active core for catalysis, as also supported by recent single-molecule studies.22 Moreover, mutational studies on the Nuc β-sheet region reported compromised TS cleavage activity.[2,40] Taken together, Nuc is critical in stabilizing the DNA TS at the level of the RuvC catalytic pocket, aiding its accommodation for cleavage, as demonstrated by our long timescale simulations.

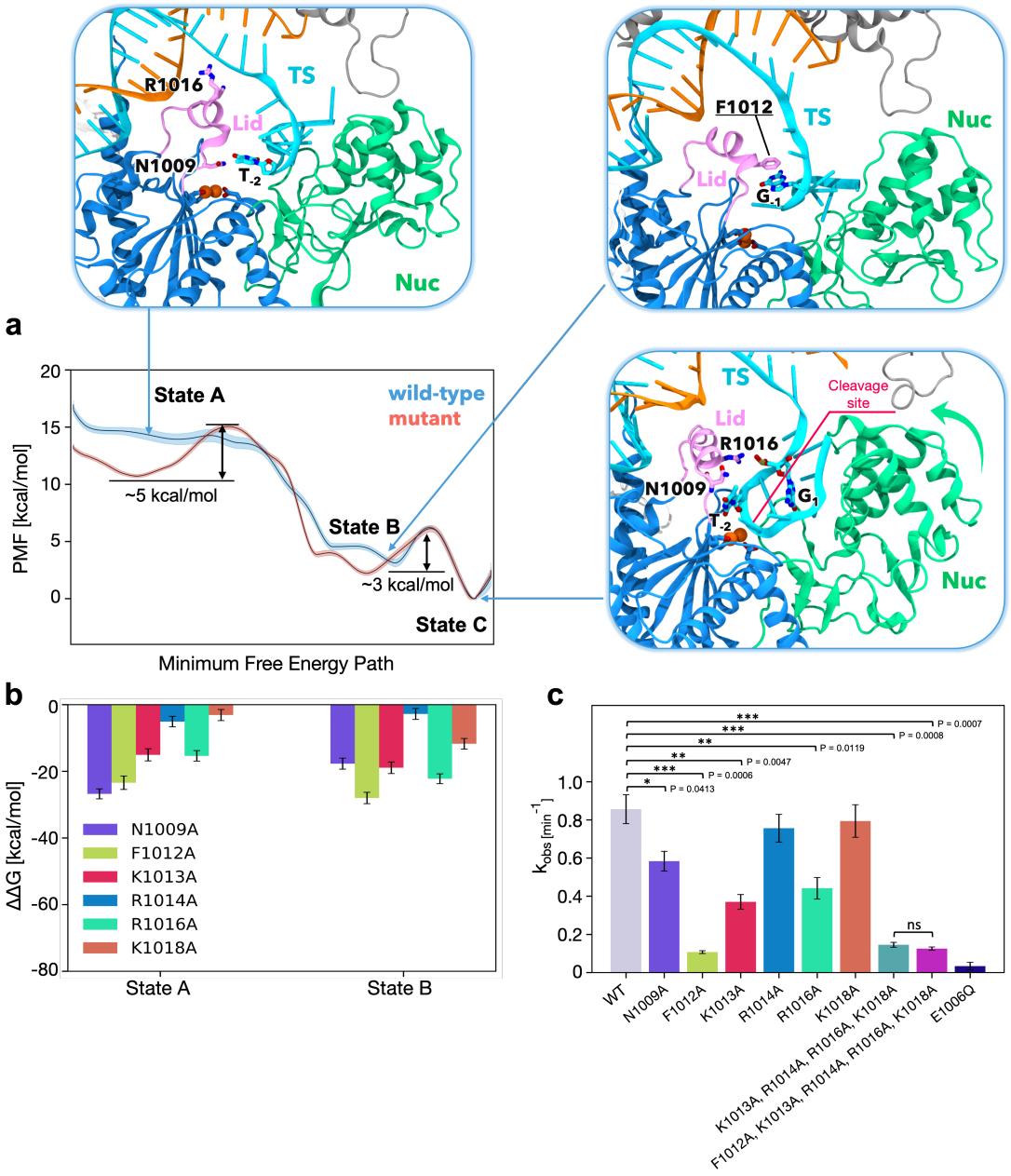

a
R1016
TS
Lid
N1009
T-2
Nuc
F1012
TS
Nuc
Lid
G-1
TS
Cleavage site
Lid
R1016
N1009
G1
T-2
Nuc
State A
wild-type
mutant
~5 kcal/mol
State B
~3 kcal/mol
State C
PMF [kcal/mol]
Minimum Free Energy Path
b
ΔΔG [kcal/mol]
N1009A
F1012A
K1013A
R1014A
R1016A
K1018A
State A
State B
c
kobs [min-1]
P = 0.0007
P = 0.0008
P = 0.0119
P = 0.0047
P = 0.0006
P = 0.0413
ns
WT
N1009A
F1012A
K1013A
R1014A
R1016A
K1018A
K1013A, R1014A, R1016A, K1018A
F1012A, K1013A, R1014A, R1016A, K1018A
E1006Q

Figure 3.4: **Traversal of the DNA target strand toward the RuvC catalytic core and role of the α-helical lid.** **(a)** Free energy profiles for the traversal of the DNA target strand (TS) toward the RuvC core, in the wild-type (WT) Cas12a (blue) and upon mutating relevant residues of the α-helical lid (N1009, F1012, K1013, R1014, R1016, K1018) into alanine (red). The Potential of Mean Force (PMF, in kcal/mol) is computed from 2-D umbrella sampling simulations and plotted along the minimum free energy path (see the Methods section). Representative snapshots, indicated by arrows, are based on the reaction coordinate values along the minimum free energy path of the wild-type Cas12a. Details for the Cas12a mutants are reported in Supplementary Figure B.10. **(b)** Relative free energy changes ($\Delta\Delta G$, in kcal/mol) of the individual residues with respect to alanine in binding the DNA TS at both the states encompassing the traversal of the TS guided by the α-helical lid, i.e., states A and B. Binding free energies are computed through the Multistate Bennett Acceptance Ratio (MBAR) method [118,119] (see the Methods section) and denoted with the associated MBAR errors. **(c)** Quantitation of *in vitro* DNA cleavage rates (k_{obs}) of Cas12a show F1012A and the combined mutation (K1013A, R1014A, R1016A, and K1018A) display a markedly significant reduction in TS cleavage, while K1013A and R1016A also significantly affect cleavage. Data represents single exponential fitting of n = 3 independent replicas and is plotted with the mean ± standard error of mean. Statistical significance was calculated using two-tailed unpaired t-test (P-value reference: ns $P > 0.05$, * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$).

3.3.4 Mutations of the Lid Adversely Affect Target Strand Traversal

To further verify the role of the lid in the traversal of the TS, we mutated the positively charged residues in the lid (K1013, R1014, R1016, K1018), along with F1012 and N1009, into alanine and we performed free energy simulations. The free energy profile was generated by collecting an ∼50 μs ensemble, following the same US approach of the wild-type (WT) system.

The PMF of the mutated system shows an additional minimum at the level of state A with a free energy barrier of ∼5 kcal/mol, which impedes a facile traversal of the TS (Figure 3.4a, red line) as it was observed in the WT system. This is because, in the mutated system, the TS locates more closely to the REC2 domain owing to the failure of interactions with the mutated lid (Supplementary Figure B.10c). The PMF of the

mutated system also shows a second free energy barrier of ∼3 kcal/mol corresponding to the TS repositioning within the RuvC catalytic core and the closure of Nuc, similar to the WT system (Supplementary Figure B.10). This indicates that alanine mutations of the lid mainly affect the traversal of the TS, with a negligible impact on the motions of major domains like REC2 and Nuc (Supplementary Figure B.15a-d). The lid loses its α-helicity briefly and most likely owing to alanine mutations,[120] the α helicity is then restored to a stable value (Supplementary Figure B.14a,d) through the length of the minimum free energy path.

The role of the aforementioned lid residues (N1009, F1012, K1013, R1014, R1016 and K1018) in aiding the traversal and accommodation of the TS in the RuvC catalytic core was further investigated by performing alchemical free energy simulations, using thermodynamic integration with softcore potentials (details are reported in the Methods section).[121] We computed the relative free energy changes (i.e., $\Delta\Delta G$) of the individual residues with respect to alanine in binding the TS at both the states encompassing the traversal of the TS guided by the α-helical lid, i.e., states A and B (Figure 3.4b). In state A, during the traversal of the TS chaperoned by the lid, N1009 and F1012 are the major contributors in binding the TS. K1013 and R1016 also play critical roles by interacting with the backbone of the TS, as also observed during classical simulations (Figure 3.2). In state B, when the TS has reached the RuvC catalytic core and is primed toward further readjustments for its accommodation, F1012 stands out as the key player in the process, while N1009, K1013, and R1016 continue to contribute in maintaining the bound state.

To experimentally verify our observations, we introduced point mutations of residues in Cas12a that were identified through molecular simulations to facilitate TS traversal

and assayed the TS cleavage activities of the mutant proteins in vitro (Figure 3.4c), where the E1006Q variant is the catalytically inactive Cas12a treated as the negative control. We observed significant reductions in TS cleavage with individual substitutions of the positively charged residues such as K1013A and R1016A, while moderate reductions were observed for point mutations of N1009A, R1014A and K1018A (Figure 3.4c, Supplementary Figure B.17). This is in accordance with alchemical free energy simulations where the free energies of K1013 and R1016 are significantly higher than R1014 and K1018 at both states A and B (Figure 3.4b). A substantial reduction in TS cleavage activity was also observed with combined alanine substitutions of all positively charged residues (K1013A, R1014A, R1016A, K1018A) in the lid, most likely due to an additive impact from K1013A and R1016A. In light of our simulations, this can be attributed to the loss of anchoring interactions between the lid and the crRNA:TS duplex (Figure 3.2). Notably, we also observed a significant reduction in TS cleavage with a single phenylalanine-to-alanine substitution (F1012A), as well as when combined with the other positively charged residues in the lid. The phenyl ring of F1012 was observed to stack with the nucleobases during umbrella sampling simulations of the TS traversal (Figure 3.4a) and displays a marked energetic contribution in state B (Figure 3.4b), suggesting that F1012 plays a critical role in guiding the TS toward the RuvC core. Taken together, our computational simulations and experimental validation establish a pivotal role of the α-helical lid in the traversal of the TS toward the RuvC catalytic pocket.

3.4 Discussion

In the CRISPR-Cas12a genome-editing system, the presence of a single RuvC catalytic site raises the question on how the enzyme could sequentially perform cleavages of the DNA NTS and TS respectively.[1,2,42] Structures of Cas12a show that the TS stays spatially distant from the RuvC catalytic core, close to the REC lobe (Figure 3.1). This implies that, following the NTS cleavage, a conformational change would allow the accommodation of the TS within the RuvC catalytic pocket, but the molecular details of this process have remained elusive.

Here, we performed extensive MD simulations of Cas12a after NTS cleavage, collecting an overall ensemble of $\sim$170 μs. The simulations reveal a mechanism for the traversal of the DNA TS toward the RuvC catalytic core, in which an α-helical lid plays a pivotal role. Structural studies showed that this α-helical lid occludes the RuvC catalytic core in the binary RNA-bound Cas12a complex and opens for the accommodation of the NTS within the catalytic site.[1,122] Multi-μs long MD simulations of Cas12a after NTS cleavage show that the α-helical lid establishes extensive interactions with the crRNA:TS duplex in its open conformation (Figure 3.2a-b). These interactions involve the positively charged residues, including K1013 and R1016, that anchor the PAM-distal end of the crRNA:TS duplex. A bending of the TS at the end of the duplex is also observed, followed by the TS region including the scissile phosphate moving toward the RuvC site (Figure 3.2c-d). This observation is in agreement with experimental findings of a distortion of the DNA TS, downstream of the crRNA:TS duplex, which was indicated to be necessary for it to reach the RuvC cleavage site.[58]

The REC2 domain also displays significant flexibility.[1,53,113] Indeed, the REC2b domain, composed of short alpha helices connected by turns, shows prominent inward rearrangement, which occurs in concert with the arching of the terminal major groove of the crRNA:TS duplex (Figure 3.3a). This suggests that the large REC2b domain aids in the TS traversal by pushing the PAM-distal major groove of the duplex toward the inner core of the protein. Both conformational changes occur concertedly after the binding of the α-helical lid to the duplex (Figures 2.2-3.3, and Supplementary Figure B.6). This indicates that the binding of the α-helical lid prompts the arching of the duplex, induced by REC2b pushing inward. Thus, the cooperativity between REC2 and the α-helical lid could foster the traversal of the DNA TS toward the RuvC core. The Nuc domain also bends inward, in concert with the bending of the DNA TS (Figure 3.3b). As REC2b and Nuc bend, they also significantly increase their interactions along the long timescale dynamics (Supplementary Figure B.7c), resulting in a closed conformation of the complex. In this respect, cryo-EM and single-molecule experiments revealed high flexibility and coupled dynamics of REC2 and Nuc,[1] also suggesting an open-to-close conformational change of the two domains.[47,52,112] Single-molecule experiments further indicated such conformational changes to be crucial for DNA cleavage.[1,53,55] Considering this evidence, the synergistic conformational changes of the major domains observed here, along with the anchoring of the α-helical lid, poise the TS to traverse toward the RuvC active site for cleavage.

To elucidate the mechanism of DNA TS traversal toward the RuvC catalytic core, we performed free energy simulations. We found that the traversal of the TS occurs barrierless toward a first free energy minimum (Figure 3.4a). This free-flowing profile

is fostered by the interactions of the α-helical lid with the duplex. Indeed, positively charged residues of the lid establish extensive electrostatic interactions with the TS, while F1012 stacks the nucleobases aiding the TS traversal. In this respect, it is interesting to note that, though we found proactive interaction of the lid with the TS in both the $\sim$10 μs-long unbiased runs (Figure 3.2) and the umbrella sampling simulations (Figure 3.4), yet the classical MD runs could not capture the barrierless transitioning of the TS toward the RuvC active site as seen from state A to B (Figure 3.4). This barrierless transition involves an $\sim$18 Å reduction between the P_{SCI} and the RuvC core, whereas classical MD could only capture the initial $\sim$6 Å reduction. Free energy simulations also show that mutations of the residues in the lid interacting with the duplex impede a facile traversal of the TS with a $\sim$5 kcal/mol barrier. Calculation of the relative free energy changes of the individual residues with respect to alanine in binding the TS at critical states of its traversal (Figure 3.4b) confirmed the pivotal role of the lid residues, especially spotlighting the contribution of F1012 in binding the TS. Our biochemical experiments also showed that alanine mutations of the interacting residues adversely impact the cleavage of the DNA TS (Figure 3.4c). While the combined mutations of the positively charged residues in the lid (K1013A, R1014A, R1016A, K1018A) result in a notable reduction of TS cleavage, single point mutation of F1012 alone plummets TS cleavage manifold times (Supplementary Figures B.17 and B.18). K1013A, and R1016A also reduce DNA cleavage, in line with binding free energy simulations. Finally, the relative free energy changes for N1009 drop moving from state A to B, which is reflected by a more moderate reduction of DNA cleavage, compared to F1012A, K1013A, and R1016A. Taken together, the lid interactions observed during our simulations establish

a pivotal role of the α-helical lid that guides the TS toward the RuvC core. In line with these observations, truncation of the lid or perturbation of its adjacent helix 1 region have also been shown to hamper target DNA cleavages.[1,123] Moreover, a structural study on CasΦ, a miniature Cas protein phylogenetically related to Cas12a, revealed a similar α-helical lid exposing the RuvC domain for cleavage.[124] Alanine substitution of the lid residues in this CasΦ protein abolished double-stranded DNA cleavages, thereby underscoring the importance of the lid in the functioning of the enzyme.

Free-energy simulations also reveal that the Nuc domain is critical in aiding the accommodation of the TS within the RuvC core. Indeed, a ∼3 kcal/mol barrier corresponds to the closure of the Nuc domain and further repositioning of the TS within the core of the enzyme (Figure 3.4a). This region of Nuc maintains persistent interactions with the 5'-tail of the DNA TS and mutational studies of this region adversely impact TS cleavage.[2,40] This suggests that Nuc could hold the TS at the RuvC site for cleavage, as also hinted by single-molecule experiments.[53] Furthermore, free energy simulations of Cas12a including a mutated lid show that the closure of Nuc remains unaffected, and mutations in the lid rather hamper only the traversal of the DNA TS. Therefore, our findings indicate that while the lid is critical for the traversal of the DNA TS, Nuc can be decisive in the accommodation and repositioning of the TS in the RuvC catalytic core.

Taking together our findings and the available cryo-EM, biochemical and single-molecule experimental data, we propose a model for the traversal of the DNA TS strand toward the RuvC core for cleavage (Figure 3.5). In this model, the α-helical lid holds a critical role, along with the conformational changes in the major domains. Upon NTS cleavage, the α-helical lid remains in its open conformation and anchors

the crRNA:TS duplex, establishing electrostatic interactions (Figure 3.5a-b). Then, the REC2 domain pushes the terminal major groove of the duplex through its REC2b region moving inward (Figure 3.5c). As REC2b pushes the crRNA:TS duplex, the Nuc domain simultaneously bends toward the core of the enzyme, aiding the bending of the DNA TS. These cooperative motions of the major domains prime the DNA TS for its traversal. The α-helical lid, now being in close proximity to the TS region around the scissile phosphate, establishes elaborate interactions with the TS and guides it toward the RuvC catalytic core. Finally, the closure of Nuc aids the TS to reposition and accommodate within the active site (Figure 3.5d).

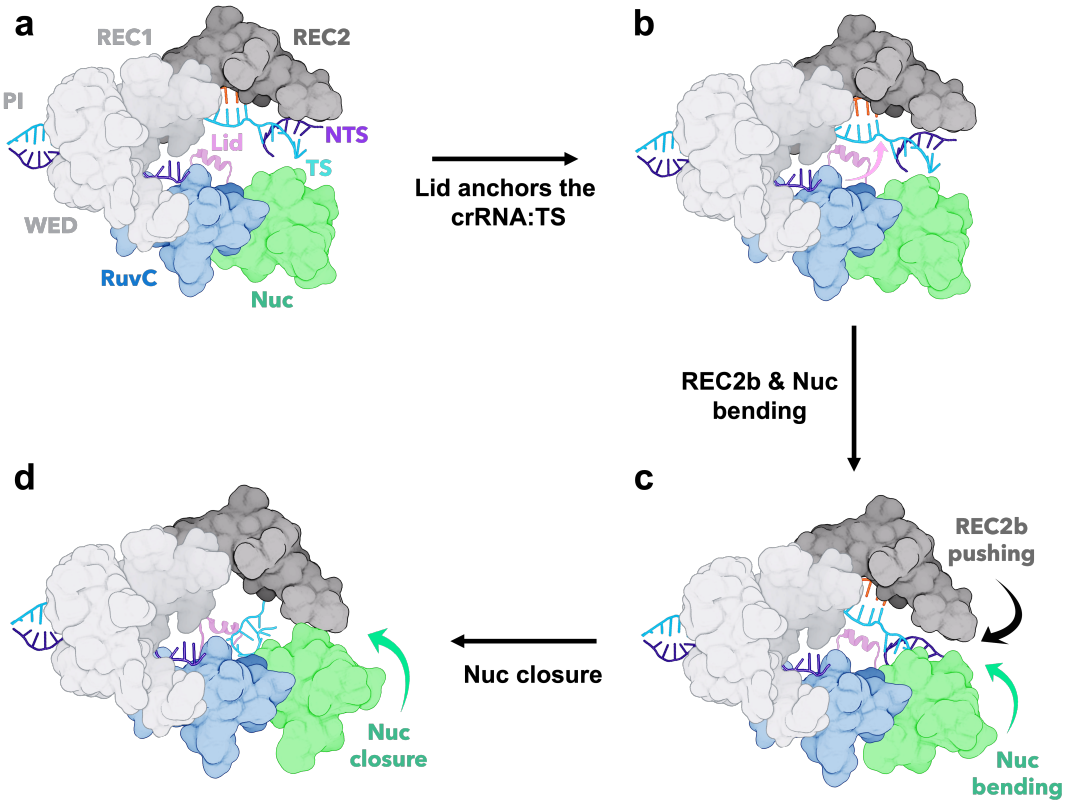


Figure 3.5: **Model for the DNA target strand traversal toward the RuvC catalytic core in CRISPR-Cas12a.** **(a)** Upon cleavage of the DNA non-target strand (NTS), the α-helical lid assumes an open conformation, exposing the RuvC catalytic core, while the DNA target strand (TS) locates close to the REC2 domain.[1] **(b)** The α-helical lid, in its open conformation, anchors the crRNA:TS duplex by establishing electrostatic interactions. **(c)** The inward bending of the REC2 and Nuc domains pushes the crRNA:TS duplex, while aiding the bending of the DNA TS and facilitating its conformational transition. **(d)** The α-helical lid binds the DNA TS and guides it toward the RuvC catalytic core. The closure of Nuc aids the TS to accommodate within RuvC for cleavage.

3.5 Materials & Methods

3.5.1 Structural Models

Molecular simulations were based on two structures of the *F. novicida* Cas12a obtained upon NTS cleavage: (1) the cryo-EM structure EMD-0065 (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation))[1] at 3.27 Å resolution and (2) the X-ray structure PDB: 5NFV [https://doi.org/10.2210/pdb6GTG/pdb] (FnCas12a bound to R-loop),[2] at 2.50 Å resolution. The 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) structure contains an α-helical "lid" (residues L1008 – K1021, Figure 3.1a), which is structurally disordered in the 5NFV [https://doi.org/10.2210/pdb6GTG/pdb] (FnCas12a bound to R-loop) structure. Hence, missing residues in the 5NFV [https://doi.org/10.2210/pdb6GTG/pdb] (FnCas12a bound to R-loop) structure were reconstructed through homology modeling using the SWISS-MODEL software (Figure 3.1b).[125] A third and more complete model system of Cas12a after NTS cleavage (Figure 3.1c) was built starting from the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) structure. Since this structure lacks the region of the TS downstream of the crRNA:TS duplex, we modeled this region of the DNA based on the 5NFV [https://doi.org/10.2210/pdb6GTG/pdb] (FnCas12a bound to R-loop) X-ray structure. The catalytically competent state of Cas12a for TS cleavage was also considered, based on related crystal structures holding the DNA TS within the RuvC active site (see Supplementary Methods). All systems were embedded in explicit waters

and counterions were added to neutralize the total charge, leading to periodic cells of $\sim$138*149*167 $\mathring{A}^3$ and $\sim$307,000 atoms for each system.

3.5.2 Molecular Dynamics (MD) Simulations

MD simulations were performed using the Amber ff19SB[126] force field, with the ff99bsc1 corrections for DNA[127] and the ff99bsc0+χOL3 corrections for RNA.[101,102] The TIP3P model was employed for water,[104] and the Li & Merz 12-6 model of non-bonded interactions was used for Mg^{2+} ions.[128] We have extensively employed these force field models in computational studies of CRISPR-Cas systems,[129] showing that they perform well for long timescale simulations on Anton-2.[75] The Li & Merz model also reported a good description of Mg^{2+} bound sites, in agreement with quantum/classical simulations.[130] An integration time step of 2 fs was employed. All bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm. Temperature control (300 K) was performed via Langevin dynamics,[107] with a collision frequency γ = 1/ps. Pressure control was accomplished by coupling the system to a Berendsen barostat[108] at a reference pressure of 1 atm and with a relaxation time of 2 ps. The systems were subjected to energy minimization, thermalization, and equilibration as detailed in the Supplementary Methods. Then, $\sim$120 ns of conventional MD were carried out in an NVT ensemble using the GPU-empowered version of AMBER 20.[131] The well-equilibrated systems were used as starting points for simulations on Anton-2,[114] a special-purpose supercomputer for multi-μs long MD simulations.

Long timescale MD simulations on Anton-2 were performed using the same force field parameters used for conventional MD simulations. A reversible multiple time step

algorithm[132] was employed to integrate the equations of motion with a time step of 2 fs for short-range non-bonded and bonded forces and 6 fs for the long-range non-bonded forces. Simulations were performed in the NPT ensemble using the multigrator integrator as implemented in Anton-2.[133] Pressure control was accomplished via the Martyna, Tobias, Klein (MTK) barostat,[134] set to maintain 1 bar of pressure, with a tau (piston time constant) parameter of 0.0416667 ps and reference temperature of 310.15 K. The barostat period was set to the default value of 480 ps per timestep. Temperature control was accomplished via the Nosé-Hoover thermostat[135,136] with a tau (time constant) parameter of 0.0416667 ps. The k-Gaussian split Ewald method[137] was used for long-range electrostatic interactions. Hydrogen atoms were added assuming standard bond lengths and were constrained to their equilibrium position with the SHAKE algorithm.[138] By using this approach, each system was simulated on Anton-2, obtaining continuous MD runs of $\sim$10 μs and in replicates. We obtained multiple trajectories of the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) cryo-EM structure (simulated for $\sim$10 μs) and its more complete model including a longer DNA ($\sim$10 μs in two replicates), as well as of the 5NFV [https://doi.org/10.2210/pdb6GTG/pdb] (FnCas12a bound to R-loop) X-ray structure ($\sim$10 μs in two replicates). The structural model of the catalytically competent Cas12a including the DNA TS within the RuvC site was also simulated for $\sim$10 μs. This resulted in a total of $\sim$60 μs of all-atom MD simulations.

3.5.3 Umbrella Sampling Simulations

The umbrella sampling (US) method[117] was used to compute the free energy profile associated with the traversal of the DNA TS toward the RuvC active site. In this method, a number of simulations (US windows) are run in parallel with additional harmonic bias potential applied to selected Reaction Coordinates (RCs):

$$V(RC) = \frac{k}{2}(RC(t) - RC^*)^2 \tag{3.1}$$

where $V(RC)$ is the value of the bias potential, k is a bias force constant, $RC(t)$ is the value of RC at given time t and RC^* is the reference value of RC. By using different RC^* values in each US window, one can sample the biased probability distribution $p_b(RC)$ along the whole RC range of interest. We performed two-dimensional (2-D) US simulations using two RCs. $RC1$ was the difference in root-mean-square deviation (RMSD) of the Cα atoms between the initial (Figure 3.1c) and the final (Supplementary Figure B.8c) states. $RC2$ was the distance between the center of mass of the RuvC catalytic core (D917, E1006, D1255) and the TS region including the scissile phosphate (i.e., the DNA nucleobases at positions -2 to -4, as indicated by biochemical and single-molecule experiments).[53,58] $RC1$ was divided into bin sizes of 0.2 Å RMSD difference and each window was run for ∼10 ns with a collective force constant k of 50 kcal mol^{-1} Å^{-2}. $RC2$ was discretized into bin sizes of 0.2 Å and individual windows were run for ∼15 ns with a collective force constant k was 50 kcal mol^{-1} Å^{-2}. Overall, we simulated 28 ($RC1$) * 119 ($RC2$) = 3,332 windows, resulting in a collective ensemble of ∼50 μs simulations. Two independent sets of 2-D US simulations were performed at 300 K

(each for $\sim$50 μs): (i) for the wild-type system and (ii) upon mutating relevant residues of the α-helical lid (N1009, F1012, K1013, R1014, R1016, K1018) into alanine. Then, the free energy profiles were computed using the Weighted Histogram Analysis (WHAM) method[139] upon removing the initial 1/3rd of the US trajectories from each window, corresponding to the relaxation of the system. Analysis of the conformational ensembles was performed on the reweighted trajectories (see Supplementary Methods). Analysis of the free energy landscape was performed by computing the minimum free energy pathway using an approach similar to the Dijkstra algorithm.[140] In this approach, the map is sampled from the origin point, propagating to the lowest energy point available on each iteration until it reaches the target point. This enabled us to consider both the *RC*s along a minimum free energy path. The error estimation on the minimum free energy pathways was performed using the Monte Carlo bootstrap error analysis.[141] Full details on the 2-D US simulations are reported in the Supplementary Methods, including the 2-D free energy surfaces (Supplementary Figure B.11) and the convergence of the free energy profiles (Supplementary Figure B.12).

3.5.4 Alchemical Free Energy Perturbation

We computed the relative free energy changes (i.e., $\Delta\Delta G$) of the individual residues with respect to alanine in binding the DNA target strand (TS), by performing alchemical free energy simulations.[142] Here, the DNA target strand (TS, ligand) bound to the Cas12a complex (receptor) was decoupled to an unbound state following the alchemical free energy formalism using softcore (SC) potentials,[121] by gradually diminishing the electrostatic and van der Waals interactions of the TS. We used a lambda (λ) dynamic

variable to define the thermodynamic states of the system along this decoupling alchemical pathway. Thermodynamic integration (TI) simulations were performed at 21 λ values from 0 to 1, at an interval of 0.05, for the wild-type (WT) CRISPR-Cas12a and for each individual alanine mutant. The systems were minimized and equilibrated using the simulation protocol described above. The final binding free energies were computed using the Multistate Bennett Acceptance Ratio (MBAR) method[118,119] to integrate the free energies over the different lambda values.[143,144] The relative binding free energy ($\Delta\Delta G$) was calculated as follows:

$$\Delta\Delta G_r esidue = \Delta G_{WT} - \Delta G_{mutated} \tag{3.2}$$

For each system (i.e., the WT CRISPR-Cas12a, and the N1009A, F1012A, K1013A, R1014A, R1016A, K1018A mutants), we considered two states of the CRISPR-Cas12a complex during the traversal of the DNA TS (states A and B, Figure 3.4). Alchemical simulations were performed at 21 values of λ (from 0 to 1, in steps of 0.05), reaching $\sim$30 ns for each window. Convergence of the $\Delta\Delta G$ for each system is reported in Supplementary Figure B.16. Considering all the systems and states investigated here, a total of $\sim$8.8 μs sampling was reached (full details are in the Supplementary Methods).

3.5.5 FnCas12a Expression and Purification

The DNA sequence of *Francisella tularensis* subsp. novicida U112 (Fn)Cas12a (WP_003040289) was codon optimized for heterologous expression in *Escherichia coli* (E. coli) and synthesized by GeneArt (Thermo Fisher Scientific). The FnCas12a gene

was inserted into the 1B plasmid (Addgene #29653) using ligation-independent cloning (LIC), resulting in a construct carrying an N-terminal hexahistidine tag followed by a tobacco etch virus (TEV) protease cleavage site. Point mutations were introduced by Gibson assembly of the PCR-amplified vector backbone with gBlock Gene Fragments (IDT) encoding the individual mutations. The sequences of the synthetic gene, primer and gBlocks are listed in Supplementary Table B.4. Mutant FnCas12a constructs were purified as for wild type. Purification of Cas12a was done as described.[145,146] In brief, Cas12a constructs were expressed in *E. coli* BL21 Rosetta2 (DE3) cells (Novagen, Wisconsin, USA). Cells were lysed in 20 mM Tris pH 8.0, 500 mM NaCl, 5 mM imidazole, 1 μg/mL pepstatin, 200 μg/mL 4-(2-Aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF) by ultrasonication. Clarified lysate was applied to a 10 ml Ni-NTA (Sigma-Aldrich) affinity column. The column was washed with 20 mM Tris pH 8.0, 500 mM NaCl, 5 mM imidazole, and bound protein was eluted by increasing imidazole concentration to 250 mM. Eluted protein was dialysed against 20 mM HEPES pH 7.5, 250 mM KCl, 1 mM DTT, 1 mM EDTA overnight at 4 °C in the presence of TEV protease to remove the 6xHis-affinity tag. Cleaved protein was further purified using a HiTrap HP Heparin column (GE Healthcare, Illinois, USA), eluting with a linear gradient to 1.0 M KCl. Elution fractions were pooled, concentrated, and further purified by size exclusion chromatography using a Superdex 200 (16/600) column (GE Healthcare) in 20 mM HEPES-KOH pH 7.5, 500 mM KCl, 1 mM DTT yielding pure, monodisperse proteins. Aliquots were flash-frozen in liquid nitrogen and stored at -80 °C.

3.5.6 FnCas12a Nuclease Activity Assays

In vitro nuclease activity assays were conducted using purified WT or mutant FnCas12a proteins programmed with a crRNA targeting the ⌊-sequence (oMS017, IDT, HPLC purified) and a dsDNA substrate containing fluorescently labeled TS (oDS285:oDS271, Merck, HPLC purified). FnCas12a and crRNA were mixed in a ratio of 1:1.2 and incubated for 10 min at 25 °C to allow binary complex formation. Reactions were started by addition of target dsDNA (FnCas12a:dsDNA, 10:1) and incubated at 37 °C. All samples were assembled in a final reaction volume of 20 μL containing 0.5 μM (mutant) FnCas12a, 0.6 μM crRNA and 50 nM dsDNA in a final buffer of 10 mM HEPES-KOH pH 7.5, 250 mM KCl, 5 mM MgCl2, 0.5 mM DTT. Reactions were stopped at indicated time points by addition of EDTA and Proteinase K (Thermo Fisher Scientific) in final concentrations of 80 mM and 0.8 mg/mL, respectively and incubated for 15 min at 37 °C. Samples were mixed with equal volume of a 2X dPAGE loading dye (95 % formamide, 25 mM EDTA), heated to 95 °C for 5 min and resolved on a 15 % denaturing (7 M Urea) polyacrylamide gels run in 0.5 X TBE buffer. Assays were conducted in technical triplicates; indicated error bars represent the standard error of mean. Fluorescence of the ATTO532-labeled substrate and cleavage products was detected using a Typhoon FLA 9500 gel imager; the cleavage rate was quantified based on loss of uncleaved substrate DNA using ImageQuant TL v.8.2.0. Sequences of RNA and DNA oligos utilized in the nuclease assays are listed in Supplementary Table B.3.

3.6 Acknowledgements

We thank Chinmai Pindi for useful discussions. This material is based upon work supported by the National Institutes of Health (Grant No. R01GM141329, to GP) and the National Science Foundation (Grant No. CHE-2144823, to GP). MJ acknowledges support from the Swiss National Science Foundation (project grant 31003A_182567) and the Howard Hughes Medical Institute (project grant 55008735). Anton 2 computer time was provided by the Pittsburgh Supercomputing Center (PSC) through Grant R01GM116961 from the National Institutes of Health. The Anton 2 machine at PSC was generously made available by D.E. Shaw Research. This work also used Expanse at the San Diego Supercomputing Center through allocation MCB160059 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Computer time was also provided by the National Energy Research Scientific Computing Center (NERSC) under Grant No M3807.

Chapter IV

Cas12a Domain Flexibility Guides R-loop Formation and Forces RuvC Resetting

Reprinted with permission from Strohkendl I. *et al.*

4.1 Abstract

The specific nature of CRISPR-Cas12a makes it a desirable RNA-guided endonuclease for biotechnology and therapeutic applications. To understand how R-loop formation within the compact Cas12a enables target recognition and nuclease activation, we used cryo-electron microscopy to capture wild-type *Acidaminococcus sp.* Cas12a R-loop intermediates and DNA delivery into the RuvC active site. Stages of Cas12a R-loop formation—starting from a 5-bp seed—are marked by distinct REC domain arrangements.

Dramatic domain flexibility limits contacts until nearly complete R-loop formation, when the non-target strand is pulled across the RuvC nuclease and coordinated domain docking promotes efficient cleavage. Next, substantial domain movements enable target strand repositioning into the RuvC active site. Between cleavage events, the RuvC lid conformationally resets to occlude the active site, requiring re-activation. These snapshots build a structural model depicting Cas12a DNA targeting that rationalizes observed specificity and highlights mechanistic comparisons to other class 2 effectors.

4.2 Introduction

CRISPR-Cas12a has been repurposed for genome editing and biotechnological applications owing to its readily adaptable RNA-guided targeting capabilities: Cas12a specifically recognizes unique DNA sequences, cleaves double-stranded DNA targets, and non-specifically cleaves single-stranded DNA (ssDNA) in *trans*.[15,31,59,147,148] Once Cas12a assembles with and processes its crRNA, it scans DNA in search of a short, T-rich protospacer adjacent motif (PAM) to initiate target recognition (Figure 4.1a). Rapid equilibrium PAM binding leads to local DNA melting that allows the CRISPR RNA (crRNA) guide to invade the duplex DNA and form a 20-base pair (bp) R-loop with the complementary target DNA.[2,48] Cas12a R-loop formation triggers activation of a single RuvC nuclease domain in rapid succession so mismatches between the crRNA guide and target strand (TS) are kinetically discriminated against during R-loop formation. The late transition state of R-loop formation ensures high specificity in both binding and cleavage.[48]

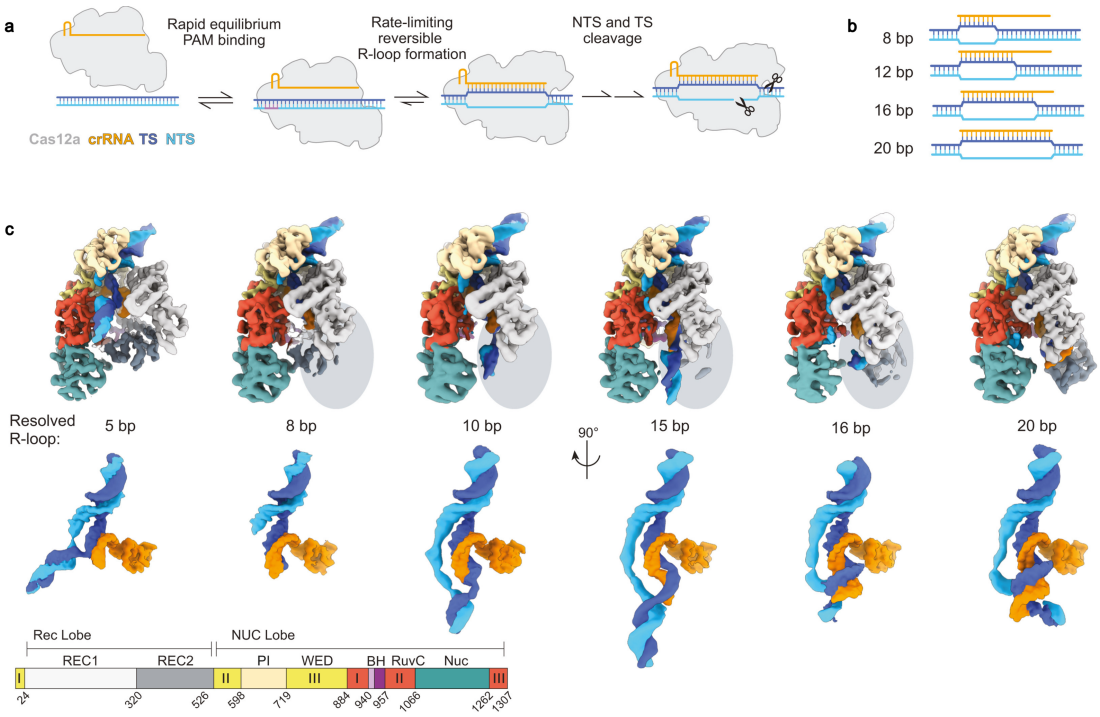


Figure 4.1: **Cas12a R-loop intermediates captured by cryo-EM. (a)** Reaction scheme describing Cas12a DNA targeting. **(b)** Target DNA substrates with varying lengths of complementarity to the crRNA guide sequence used for cryo-EM. **(c)** DNA-bound Cas12a cryo-EM structures are shown in order of increasing R-loop base pairs resolved. Below, densities for nucleic acids are shown at a 90° angle to the overall structure. Varying contour thresholds were used to produce continuous NTS density. Maps are colored according to the protein domain map and R-loop schematic. REC1 and REC2, Recognition domains; PI, PAM-interacting domain; WED, Wedge domain; BH, Bridge helix; RuvC, RuvC nuclease domain; Nuc, Nuc domain. Gray ovals are used to highlight missing density for the REC2 domain. See also Supplementary Figures C.1 and C.2.

Cas12a functions similarly to the well-studied SpyCas9 as both rely on rate-limiting R-loop formation by a 20-nt crRNA guide to activate target DNA cleavage.[36,90,149,150] However, differences in activity and specificity profiles suggest they have distinct underlying mechanisms that must be influenced by their protein architecture.[151–153] From Cas12a structures, it is known that the REC lobe must rearrange to accommodate the helical 20-bp R-loop and expose the RuvC nuclease domain.[1,2,16,41,43,45,49,72] Single molecule studies have further described protein conformational changes that enable

double stranded DNA cleavage.[42,52,53,62,123,154] Given the importance rate-limiting R-loop formation has in nuclease activation and off-target rejection, structures of the R-loop would lead to better understanding of the underlying mechanism driving Cas12a activity and specificity. Additionally, as a new wave of minimal type V CRISPR effectors are being pursued for genome editing applications, understanding the functional importance of lost domains could guide engineering efforts to improve their efficiency.[155–160]

To examine how Cas12a interrogates its DNA target and transitions into the catalytically active state, we used cryo-electron microscopy (cryo-EM) to capture structures of wild type (WT) Cas12a during R-loop formation and DNA cleavage. We show that Cas12a first interrogates 5-bp of target DNA before dramatic REC domain flexibility accommodates R-loop extension but precludes contacts that would stabilize intermediates and drive propagation. Only during formation of the final few base pairs do all domains fully dock onto the R-loop, forming contacts throughout the majority of the RNA-DNA heteroduplex. Delayed domain docking is also associated with contacts that promote RuvC exposure. The non-target strand (NTS) is then captured in the RuvC active site, poised for cleavage. After NTS cleavage, we observe dramatic movement while the downstream TS bends into the RuvC active site.

4.3 Results

4.3.1 Structures of Cas12a on Pathway to Cleavage

To capture various intermediates of WT Cas12a during R-loop formation, we used DNA duplex substrates with increasing amounts of complementarity to the crRNA

guide assembled with WT *Acidaminococcus sp.* Cas12a (herein referred to as Cas12a) (Figure 4.1b, Supplementary Table C.2). Cas12a-crRNA-DNA incubation times were varied to maximize binding and minimize cleavage before vitrification, guided by our previous kinetic characterization.[48,62] From these four cryo-EM samples, we obtained six distinct R-loop intermediate structures with nominal resolutions ranging from 3.3-3.7 Å (Figure 4.1c, Supplementary Figures C.1-C.2, Table 4.1).

The 8-bp target dataset yielded two structures with well resolved 5- and 8-bp of contiguous R-loop density (5- and 8-bp R-loops; Figure 4.1c, Supplementary Figure C.2a). An additional class of 5-bp R-loop without a resolved REC1 demonstrates the domain flexibility during R-loop seeding. Interestingly, in the 5- and 8-bp structures, density for the unpaired crRNA guide continues past the nascent R-loop, suggesting a repeated pre-ordering mechanism to promote efficient R-loop propagation to subsequent bases.

The 12-bp target dataset produced a new intermediate with only 10 bp of the R-loop resolved and an unresolved REC2 domain (Figure 4.1c, Supplementary Figure C.2b), correlating extension past the seed through the middle of the R-loop with high conformational flexibility. The 16-bp target dataset continued to highlight the dynamic nature of the REC2 domain and structural heterogeneity within the PAM-distal half of the R-loop (Supplementary Figure C.2c). Within this dataset we captured clear 15- and 16-bp structures with the NTS approaching the RuvC active site (Figure 4.1c). Other structural classes within the 16-bp dataset had unresolved REC2 domains and occluded RuvC active sites. Finally, we resolved Cas12a with the full R-loop and the intact NTS poised for cleavage in the catalytically-competent RuvC active site (Figure

4.1c, Supplementary Figure C.2e). This previously uncaptured structure demonstrates the mechanism of RuvC-mediated cleavage while remaining consistent with previously published x-ray crystal structures[2,16,41,72] and cryo-EM models.[1]

Together these structures of WT Cas12a depict transient intermediate states of R-loop and lead to several broad observations: (1) the number of R-loop base pairs resolved is not always equal to the extent of target complementarity, hinting at the relative position of energetic barriers during R-loop formation; (2) REC2 is not resolved during middle R-loop propagation; and (3) the distal DNA migrates from projecting out the front of Cas12a to the back during R-loop propagation and threads the NTS along the RuvC domain.

Table 4.1: Cryo-EM Data Collection and Model Statistics

	5-bp (EMDB-40441) (PDB 8SFH)	**8-bp** (EMDB-40442) (PDB 8SFI)	**10-bp** (EMDB-40443) (PDB 8SFJ)	**15-bp** (EMDB-40444) (PDB 8SFL)	**16-bp** (EMDB-40445) (PDB 8SFN)	**20-bp** (EMDB-40446) (PDB 8SFO)	**TSps-TS in active site** (EMDB-40447) (PDB 8SFP)	**TSps-RuvC exposed** (EMDB-40448) (PDB 8SFQ)	**TSps-Extended** (EMDB-40449) (PDB 8SFR)
Data collection and processing									
Voltage (kV)	300	300	300	300	300	200	300	300	300
Electron exposure (e–/Å^2)	80	80	80	80	80	49	80	80	80
Defocus range (μm)	-1.5 to -2.5								
Pixel size (Å)	0.833	0.833	0.833	0.833	0.833	0.94	0.833	0.833	0.833
Symmetry imposed	C1								
Initial particle images (no.)	1,624,766	1,624,766	1,509,051	3,758,114	3,758,114	582,836	3,288,205	3,288,205	3,288,205
Final particle images (no.)	60,595	47,461	122,723	158,948	97,787	154,794	25,296	104,326	105,267
Map resolution (Å)	3.42	3.51	3.35	3.27	3.29	3.33	3.76	3.45	3.57

Continued on next page

continued from previous page

	5-bp (EMDB-40441) (PDB 8SFH)	**8-bp** (EMDB-40442) (PDB 8SFI)	**10-bp** (EMDB-40443) (PDB 8SFJ)	**15-bp** (EMDB-40444) (PDB 8SFL)	**16-bp** (EMDB-40445) (PDB 8SFN)	**20-bp** (EMDB-40446) (PDB 8SFO)	**TSps-TS in active site** (EMDB-40447) (PDB 8SFP)	**TSps-RuvC exposed** (EMDB-40448) (PDB 8SFQ)	**TSps-Extended** (EMDB-40449) (PDB 8SFR)
FSC threshold	0.143								
Map resolution range (Å)	2.5 – 6.5								
Refinement									
Initial model used (PDB code)	5B43								
Model resolution (Å)	3.7	3.8	3.8	3.6	3.6	3.6	3.9	3.7	3.8
FSC threshold	0.5								
Map sharpening B factor (Å^2)	132.2	128.5	173.6	162.8	145.3	154.4	152.1	165.7	172.4
Map correlation coefficient	0.83	0.78	0.76	0.78	0.72	0.81	0.83	0.87	0.88
Model composition									

Continued on next page

continued from previous page

	5-bp (EMDB-40441) (PDB 8SFH)	**8-bp** (EMDB-40442) (PDB 8SFI)	**10-bp** (EMDB-40443) (PDB 8SFJ)	**15-bp** (EMDB-40444) (PDB 8SFL)	**16-bp** (EMDB-40445) (PDB 8SFN)	**20-bp** (EMDB-40446) (PDB 8SFO)	**TSps-TS in active site** (EMDB-40447) (PDB 8SFP)	**TSps-RuvC exposed** (EMDB-40448) (PDB 8SFQ)	**TSps-Extended** (EMDB-40449) (PDB 8SFR)
Non-hydrogen atoms	12407	11477	10207	12973	12807	12606	12062	12045	12047
Protein residues	1286	1227	1001	1296	1302	1240	1240	1240	1240
Nucleotides	91	70	97	114	104	119	93	92	92
Ligand	MG: 2								
Mean B factors (Å^2)									
Protein	72.43	83.33	75.53	82.36	86.01	64.92	127.68	121.39	109.43
Nucleotides	131.56	106.3	129.35	182.19	141.65	105.18	124.34	122.06	115.71
R.M.S. deviations									
Bond lengths (Å)	0.006	0.005	0.005	0.005	0.006	0.004	0.005	0.005	0.005
Bond angles (°)	0.991	0.986	1.059	0.912	1.028	0.798	1.095	0.994	0.948
Validation									
MolProbity score	1.05	1.44	1.41	1.48	1.45	1.27	1.56	1.39	1.63

Continued on next page

continued from previous page

	5-bp (EMDB-40441) (PDB 8SFH)	**8-bp** (EMDB-40442) (PDB 8SFI)	**10-bp** (EMDB-40443) (PDB 8SFJ)	**15-bp** (EMDB-40444) (PDB 8SFL)	**16-bp** (EMDB-40445) (PDB 8SFN)	**20-bp** (EMDB-40446) (PDB 8SFO)	**TSps-TS in active site** (EMDB-40447) (PDB 8SFP)	**TSps-RuvC exposed** (EMDB-40448) (PDB 8SFQ)	**TSps-Extended** (EMDB-40449) (PDB 8SFR)
Clashscore	2.68	5.89	6.84	5	6.47	3.81	7.51	6.61	8.12
Poor rotamers (%)	0.52	0.55	0.34	0.69	0.26	0.18	0.54	0.72	0.63
Ramachandran plot									
Favored (%)	98.51	97.37	97.88	96.59	97.53	97.49	97.16	97.89	96.84
Allowed (%)	1.49	2.63	2.12	3.41	2.47	2.51	2.84	2.11	3.16
Disallowed (%)	0	0	0	0	0	0	0	0	0

4.3.2 Cas12a Uses a 5-bp Seed

Cas12a, like other RNA-guided effectors, relies on a seed region within the R-loop to promote efficient guide:target Watson-Crick base pairing and stringent discrimination against sequence mismatches.[2,8,161–163] A previously published binary complex of Fn-Cas12a7 showed the first five nucleotides of the crRNA guide exposed and pre-ordered in A-form-like geometry, suggesting a 5-bp seed consistent with preliminary mismatch data.[15] Here, we present the first structures of Cas12a recognizing a 5-bp seed.

The 5-bp structure closely resembles Cas12a bound to guide RNA (Supplementary Figure C.3a) (Fn: 6.38 Å root-mean-square deviation [RMSD]).[2] The first three nucleotides of displaced NTS are gripped by the PAM Interacting (PI) domain. The REC1 domain has shifted outward to accommodate formation of an A-form heteroduplex. The majority of contacts are concentrated within the first few base pairs of the R-loop and are made by the Wedge (WED) and PI domains (Figure 4.2a, Supplementary Figure C.3b). The REC1 domain makes few non-specific stabilizing contacts to the R-loop at the end of the seed (K51, N175, R176). The TS and NTS rehybridize at the 6^{th} base pair positioned between the RuvC domain and the REC1 helix-loop-helix, directing the distal DNA to exit the complex. The DNA at the R-loop-DNA junction sits atop a bulky loop in the RuvC domain (previously referred to as "septum"[49]) with K1054 projecting into the minor groove of the DNA (Figure 4.2a).

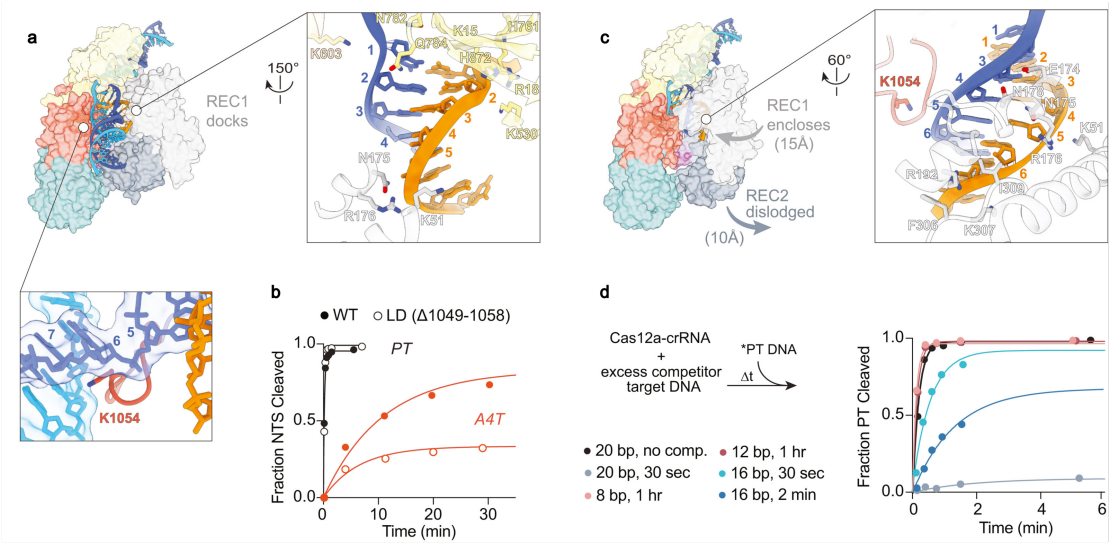


Figure 4.2: **Cas12a rapidly checks for DNA complementarity with a 5-bp seed.** **(a)** Structure of Cas12a with a 5-bp R-loop shown in surface representation. Top zoomed-in view highlights the WED and REC1 contacts of the 5-bp R-loop. Portions of REC1 that do not contact the R-loop have been removed for clarity. Bottom zoomed-in view highlights the R-loop-DNA junction that sits atop the bulky RuvC loop. **(b)** WT and LD NTS cleavage of the matched 20-bp target (PT) and a single mismatch target (A4T). WT Cas12a marked by solid circles, LD Cas12a marked by hollow circles. **(c)** Structure of Cas12a with an 8-bp R-loop shown in surface representation. Arrows denote domain movements from the 5-bp structure. Zoomed-in view of the 8-bp R-loop highlights additional REC1 contacts along the crRNA backbone and minor groove. **(d)** Competition cleavage assay to test the ability of R-loop intermediate structures to stably bind Cas12a and inhibit cleavage of a matched PT DNA. Competitor DNAs were pre-incubated with Cas12a (variable times noted) before addition of labeled PT DNA. Plot shows competition time courses of NTS cleavage of the PT substrate. In **(b)** and **(d)**, example cleavage reactions with six time points are fit to a single exponential curve (not all time points are shown). Only a single replicate is shown for each condition. See also Supplementary Figure C.3

During the processing of this dataset, we also identified a structural class with an identical 5-bp R-loop and an unresolved REC1 domain (density is nearly absent starting where REC1 hinges, Supplementary Figure C.3c). This reconstruction most likely precedes the Cas12a with the resolved REC1 domain and hints at REC1 flexibility and a static REC2 during initial target search.[42,47,164]

To probe the role of the RuvC loop within the early R-loop structures, we purified a loop deletion mutant (LD) mutant (Δ1049-1058) and measured its effect on cleavage of a matched and mismatched target (Figure 4.2b). LD did not have a measurable effect on NTS cleavage of the matched target (PT) (5.03 ±0.44 vs 6.25 ±0.73 min^{-1} for WT) but resulted in reduced cleavage of a single nucleotide mismatch (A4T) at position 4 of the R-loop (0.080 ±0.016 vs 0.172 ±0.015 min^{-1} for WT; Supplementary Tables C.3 and C.4). The minor reduction in cleavage rate of the seed mismatch substrate reports that the loop was contributing to the stability of the 5-bp intermediate over the PAM-bound state in the WT enzyme. The increase in observed specificity by the LD mutant likely arises by destabilizing the early R-loop intermediate and increasing the rate of R-loop collapse for both DNA substrates, resulting in a measurable decrease in the presence of mismatches.

Given the differences in targeting efficiency between *in vitro* and cellular contexts, we tested the effect of the seed-stabilizing loop in *E. coli* using a plasmid interference assay (Supplementary Figure C.3d).[165] WT Cas12a led to efficient plasmid interference when targeting the matched target (Supplementary Figure C.3e), however expression of the LD mutant led to minimal plasmid interference as seen by reduced GFP fluorescence. These results support the importance of the bulky loop for seed stability during R-loop formation.

4.3.3 Fast Dissociation Kinetics of Early and Middle R-loop Intermediates Promote Efficient Target Search

The 8-bp intermediate resembles previous Cas12a structures with an 8-bp R-loop (Fn: 10.6 Å RMSD; Lb: 5.3 Å RMSD)[1,54] (Supplementary Figure C.3a). The REC1 domain rotates toward the NUC lobe by ∼15 Å and encloses on the R-loop, forming a network of new contacts along the crRNA backbone from positions 5 to 8 and along the TS (Figure 4.2c, Supplementary Figure C.3f). The newly formed base pairs and contacts could be expected to strongly stabilize the 8-bp intermediate; however this would contradict the working model that Cas12a relies on a readily reversible R-loop to discriminate against incorrect DNA sequences.

To functionally measure the stability of shorter R-loop species, we tested their ability to competitively inhibit Cas12a targeting the matched target DNA (Figure 4.2d). The 8- or 12-bp DNA targets pre-bound to Cas12a (pink and brown, respectively) did not result in reduced cleavage of the PT DNA (black)—neither in amplitude nor rate—reporting that Cas12a R-loops 12-bp or shorter rapidly collapse at a rate faster than R-loop formation. This is consistent with the lack of observed single-molecule binding events to short targets.[42] The 16-bp competitor target (cyan and dark blue) demonstrated inhibition dependent on pre-incubation time, reporting that this intermediate R-loop can be stably bound but likely remains within the readily reversible region of the R-loop. The 20-bp competitor effectively inhibits all cleavage of the PT DNA and demonstrates that Cas12a is stably bound to a matched target by 30 s with little dissociation. This rapid R-loop intermediate collapse likely acts as a mechanism to ensure Cas12a does

not get sequestered at incorrect DNA sequences and is able to efficiently search for the correct target.

The fast dissociation kinetics of these competition reactions means the early and middle intermediates observed in our cryo-EM datasets represent transient structures during reversible R-loop formation. Indeed, reconstructions resembling the 5- and 8-bp R-loop structures were also identified in the datasets of the 12- and 16-bp targets, further supporting that these states are representative of early intermediates on-pathway to DNA cleavage.

4.3.4 REC2 Flexibility Accommodates R-loop Extension but Delays Contacts

REC2 domain conformational flexibility is one of the most dramatic series of changes seen within the R-loop intermediate structures. A structural consequence of REC2 domain flexibility is the structural state of the BH domain (here considered inclusive of RuvC II helix 1). Our R-loop intermediate structures show that the BH W958 starts to dislodge from the increasingly mobile REC2 in the 8-bp structure, resulting in BH flexibility during middle R-loop propagation (Figure 4.3a). This lack of rigidity allows a necessary conformational rearrangement within the two BH helices but prevents contacts from stabilizing nascent R-loop base pairs as they form (Figure 4.3b). In the 16-bp structure, REC2 begins to dock onto the R-loop in the correct orientation, but remains at low occupancy likely due to insufficient stabilizing contacts. Additionally, the BH is anchored within REC2 but the helices have not yet fully formed (Figure 4.3a). REC2 domain flexibility during middle and late R-loop formation is further demonstrated in

the remaining structural classes from the 12- and 16-bp datasets that have an unresolved REC2 domain (Supplementary Figures C.2b and C.2c).

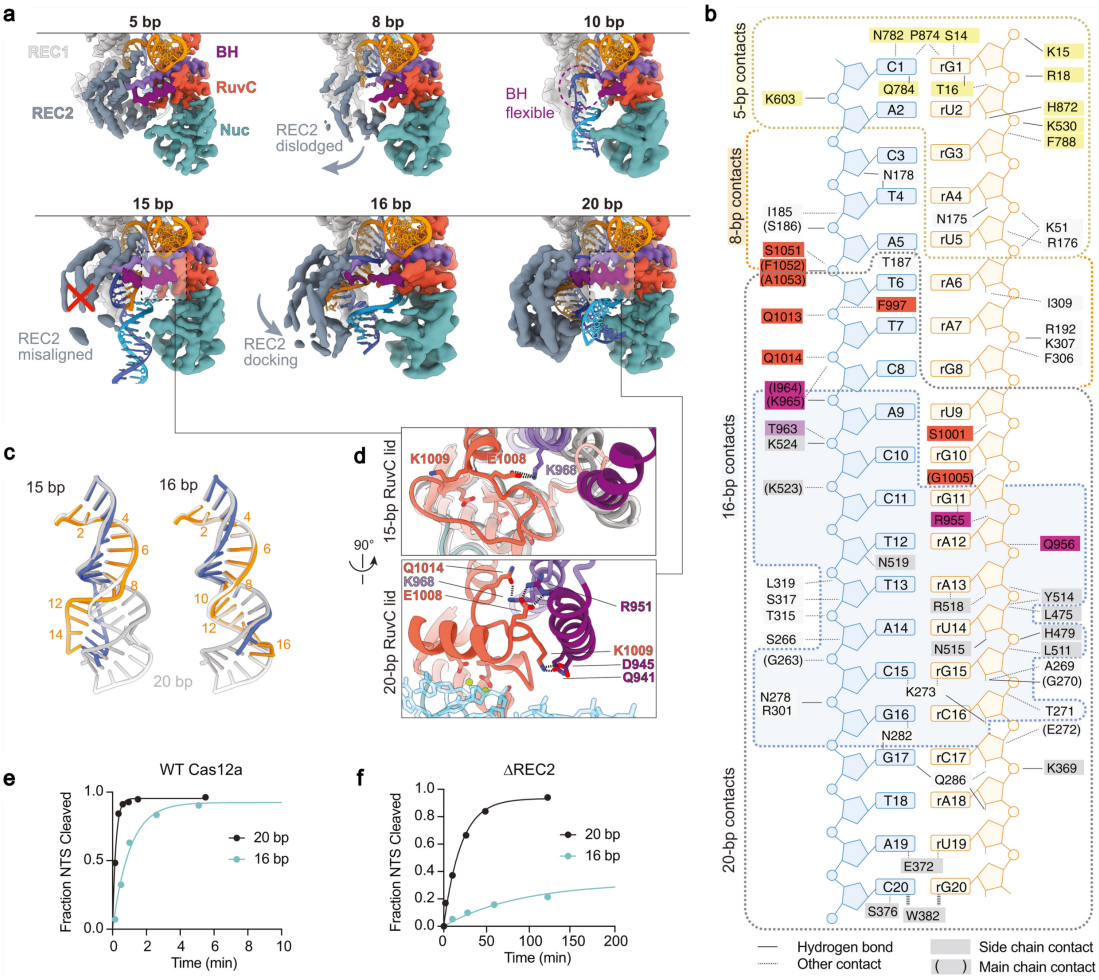

a
5 bp
8 bp
10 bp
REC1
BH
RuvC
REC2
Nuc
REC2 dislodged
BH flexible
15 bp
16 bp
20 bp
REC2 misaligned
REC2 docking
b
5-bp contacts
8-bp contacts
16-bp contacts
20-bp contacts
N782 P874 S14
K15
C1
rG1
R18
Q784
T16
K603
A2
rU2
H872
K530
F788
C3
rG3
N178
T4
rA4
I185 (S186)
N175
K51
R176
S1051 (F1052) (A1053)
A5
rU5
T187
T6
rA6
I309
F997
Q1013
T7
rA7
R192
K307
F306
Q1014
C8
rG8
(I964) (K965)
A9
rU9
T963 K524
S1001
C10
rG10
(K523)
(G1005)
C11
rG11
R955
T12
rA12
Q956
N519
L319
T13
rA13
Y514
S317
R518
L475
T315
A14
rU14
H479
S266
N515
L511
(G263)
C15
rG15
A269 (G270)
K273
N278 R301
G16
rC16
T271 (E272)
N282
G17
rC17
K369
Q286
T18
rA18
A19
rU19
E372
C20
rG20
S376
W382
Hydrogen bond
Other contact
Side chain contact
( ) Main chain contact
c
15 bp
16 bp
20 bp
d
15-bp RuvC lid
20-bp RuvC lid
90°
K1009
E1008
K968
Q1014
K968
E1008
R951
K1009
D945
Q941
e
WT Cas12a
Fraction NTS Cleaved
Time (min)
20 bp
16 bp
f
ΔREC2
Fraction NTS Cleaved
Time (min)
20 bp
16 bp

Figure 4.3: **Middle R-loop propagation is marked by REC2 flexibility, delaying R-loop contacts and RuvC activation.** **(a)** Structures of Cas12a showing the lower back half of the protein during R-loop propagation. Protein domains are shown as the reconstruction maps and nucleic acids are shown as models. WED and PI domains are removed for clarity. REC2 becomes dislodged as the R-loop extends past the seed and then returns to a cleavage-competent conformation after R-loop reaches 16 bp. BH domain flexibility is also seen. **(b)** Diagram of the 20-bp R-loop contacts grouped according to length of the R-loop when contacts were resolved (5-bp: yellow, 8-bp: orange; 16-bp: blue; 20-bp: gray). R-loop and amino acids are colored as in Figure 4.1. BH contacts from the 16-bp structure are associated with less-well resolved signal and called with reduced confidence. **(c)** Overlay of late R-loop intermediates onto the 20-bp R-loop (gray) to demonstrate the change in R-loop shape. crRNA guide bases are labeled. **(d)** Zoomed-in side view of RuvC lid and BH highlights the change in contacts made by E1008 and K1009 during R-loop formation. The 15-bp model is overlaid with the 5-, 8-, and 10-bp intermediates in gray. See also Supplementary Figure C.4. **(e)** Example time course of NTS cleavage by Cas12a of DNA substrates that form a 16- or 20-bp R-loop. **(f)** Example time course of NTS cleavage by the ΔREC2 mutant.

Additionally, the growing R-loop is centrally located within the Cas12a channel and diverges from the path of the final 20-bp R-loop heteroduplex (Figure 4.3c). This contrasts with the first half of the R-loop, which remains in an A-form helix. As captured in the 15-bp structure, starting at position 10, the minor groove narrows and the heteroduplex stretches to enable TS:crRNA base pairing despite the constraints imposed by the continuous NTS lining the NUC lobe. The resulting deformed and mispositioned distal R-loop prevents premature REC2 docking into the correct arrangement.

Only after the 16^{th} base pair is formed, domain docking is observed and upstream contacts are made throughout the R-loop via the backbone and minor groove (Figure 4.3b, Supplementary Figure C.3g): REC1 shifts inward to the R-loop contacting the TS from positions 13 to 16; REC2 begins to dock within the minor groove, contacting both phosphate backbones and several bases from position 10 to 14; and the BH begins to make contact with the crRNA backbone at positions 11 and 12. Finally, formation of the complete 20-bp R-loop supports stable REC2 (and BH) docking and additional contacts are made throughout, from positions 8 to 20 of the R-loop.

The observed flexibility of Cas12a and the structurally heterogeneous nature of the middle R-loop intermediates show that propagation past the early R-loop proceeds without protein contacts guiding heteroduplex formation. Instead, R-loop propagation must be driven by the energetics of DNA base pair melting and RNA:DNA base pair formation. The lack of stabilizing contacts by Cas12a is also a logical explanation for the rapid R-loop collapse of early and middle R-loop intermediates.

4.3.5 Late REC2 Docking Promotes Catalysis

Cas12a requires an R-loop of at least ∼15 bp to activate cleavage by the RuvC nuclease.[1,15,62,63] The integrity of the BH has also been shown to influence DNA cleavage and offers a method for signal transduction—communicating sufficient R-loop recognition by the REC lobe to the NUC lobe where the RuvC domain is located.[52,123,166] Our late R-loop intermediates provide the structural framework to explain how the minimum R-loop length leads to nuclease activation.

As the R-loop forms, the RuvC active site is occluded by the RuvC lid up to the 15-bp structure. At the same time, changes to the BH are seen during R-loop formation: by 10 bp the BH is partially unresolved and at 15 bp the BH can be kinked away from the R-loop due to a misaligned REC2 domain (Figure 4.3d). In contrast, the 20-bp R-loop with the RuvC active site exposed to a coordinated NTS has the BH anchored into a properly docked REC2 and contacts the minor groove of the R-loop. The conformational rearrangement of the BH enables contacts with the newly formed RuvC lid α-helix that now exposes the active site: lid K1009 contacts the BH via Q941 and D945, and lid E1008 is pulled up by BH R951, supporting the raised conformation of the lid (Figure 4.3d). These structural changes are consistent with contacts seen in the binary and post-cleavage ternary complex of FnCas12a.[1,2,49]

The coordinated changes seen in the early R-loops versus the complete 20-bp R-loop suggest that REC2 docking contributes to RuvC exposure via the BH. Accordingly, unstable REC2 docking in the 16-bp structure allows flexibility in the BH and RuvC lid α-helix that reduces contacts and causes less-well resolved density (Supplementary Figure C.4a). Our structural observations are consistent with previous biochemical

characterizations of the BH, which demonstrate that structural integrity of the helices and the anchoring tryptophan is important for DNA cleavage.[45,52,123,166,167] A proposed mechanistic model resulting from this combined data is that R-loops with sufficient lengths to permit semi-stable REC2 docking will achieve RuvC activation via BH contacts but at reduced efficiencies caused by REC2 and BH flexibility.

In support of this model, NTS cleavage of the 16-bp target occurred at a 5-fold reduced rate compared to that of the 20-bp R-loop (1.32 $\pm$0.31 min^{-1}; Figure 4.3e). Because rate-limiting R-loop formation of the 20-bp target could mask penalties on cleavage, we also tested *trans*-cleavage by Cas12a activated with single stranded substrates to isolate the RuvC exposure step (Supplementary Figure C.4b). The 16-bp R-loop resulted in 22-fold slower rates than the 20-bp R-loop (3.8 ($\pm$0.57) vs 81 ($\pm$11) x 104 M^{-1} min^{-1}, respectively) which is in agreement with the $\geq$5-fold penalty observed for NTS cleavage (Figure 4.3e). To further test the role of REC2 docking in RuvC activation via anchoring the BH, we removed the REC2 domain (ΔREC2). Cleavage of the 20-bp target NTS was dramatically reduced by 65-fold (0.0963 $\pm$0.0081 min^{-1}) compared with WT Cas12a and NTS cleavage of the 16-bp target was reduced 83-fold (0.0158 $\pm$0.0089 min^{-1}; Figure 4.3f and Supplementary Figure C.4c). This Cas12a mutant confirms the importance of the REC2 domain for RuvC activation and, interestingly, resembles a hypothetical ancestral Cas12 effector that lacks REC domain insertions. Similarly, these enzymes are marked by poor nuclease activity.[160]

These data highlight the role of REC2 docking in sensing R-loop length and communicating the signal to the RuvC domain via the BH, promoting activation. Because

REC2 requires that the R-loop has both sufficient length (∼16 bp) and correct geometry to properly dock, it acts as a delayed checkpoint for Cas12a activation. Off-target DNA sequences outside of the seed could favor R-loop collapse or introduce R-loop helix distortions, such as kinks or minor groove changes, limiting the likelihood of REC2 docking.[115,168] This late conformational checkpoint enables specificity against mismatches positioned late in the R-loop and, once achieved, promotes RuvC exposure.

4.3.6 NTS Poised for Cleavage in the Active Site

During final base pairing, domain docking begins and the R-loop-DNA junction moves from being positioned within the central cavity of Cas12a (16-bp) to exiting out the back of Cas12a between REC2 and the NUC domain stem (20-bp), causing the elongating single-stranded NTS to traverse across the RuvC domain, just below the lid α-helix (Figure 4.4a and Supplementary Figure C.4d). Thanks to short sample incubation time guided by previous kinetic characterization of DNA targeting,[48] the 20-bp structure captures the NTS in the RuvC active site before cleavage.

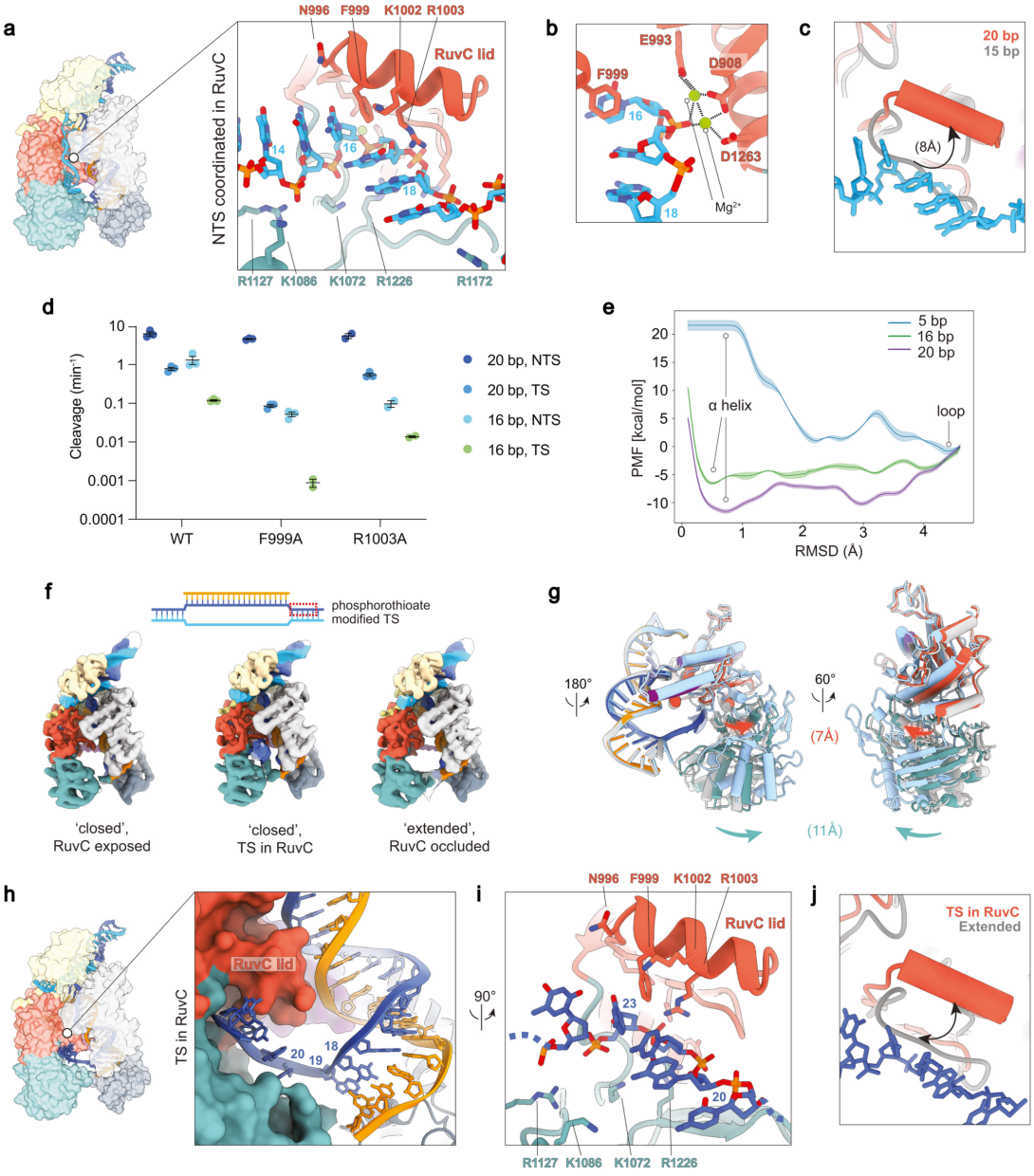

a
N996 F999 K1002 R1003
RuvC lid
NTS coordinated in RuvC
14
16
18
R1127 K1086 K1072 R1226 R1172
b
E993
F999
D908
D1263
16
18
Mg2+
c
20 bp
15 bp
(8Å)
d
Cleavage (min-1)
10
1
0.1
0.01
0.001
0.0001
WT
F999A
R1003A
20 bp, NTS
20 bp, TS
16 bp, NTS
16 bp, TS
e
PMF [kcal/mol]
20
15
10
5
0
-5
-10
0
1
2
3
4
RMSD (Å)
5 bp
16 bp
20 bp
α helix
loop
f
phosphorothioate
modified TS
'closed',
RuvC exposed
'closed',
TS in RuvC
'extended',
RuvC occluded
g
180°
60°
(7Å)
(11Å)
h
TS in RuvC
RuvC lid
20
19
18
i
90°
N996 F999 K1002 R1003
RuvC lid
23
20
R1127 K1086 K1072 R1226
j
TS in RuvC
Extended

Figure 4.4: **Structures of RuvC-mediated DNA cleavage by Cas12a.** **(a)** Structure of Cas12a with a complete 20-bp R-loop and continuous NTS in the RuvC active site shown in surface representation. Zoomed-in view of the NTS held in the RuvC active site. F999 base stacks with base 16 of the NTS upstream of the scissile phosphate. See also Supplementary Figure C.4 **(b)** Coordinated RuvC active site for NTS cleavage. Catalytic residues (D908, E993 and D1263) are labeled. **(c)** Overlay of 15- (gray) and 20-bp structures shows the steric clash between the NTS and the RuvC lid when not arranged as an α-helix. **(d)** NTS and TS cleavage of the 20- and 16-bp substrates measured for WT and RuvC lid mutants. Colored circles represent individual replicates. Black lines define the mean ± standard error of the mean (SEM). **(e)** Free energy profiles capturing the transition of the RuvC lid from α-helix (RMSD ∼1 Å) to unstructured loop (RMSD ∼4 Å) during different stages of R-loop formation (Supplementary Figure C.4h show potential mean force [PMF] profile convergence). The standard deviation of the free energy profiles obtained for the respective convergence plots is shown by the translucent colors. **(f)** Post-NTS cleavage Cas12a structures show Cas12a in a "closed" conformation with the active site exposed or occupied by the TS, and Cas12a in an "extended" conformation. See also Supplementary Figures C.1 and C.5. **(g)** Models from **(f)** aligned by REC1 and the R-loop to show extent of NUC lobe movements to bring the TS to the active site. **(h)** Structure of Cas12a with the TS in the RuvC active site shown in surface representation. Zoomed-in view of R-loop base pairs 18–20 melted to enable TS bending into the RuvC active site. **(i)** Zoomed-in view of the TS interacting with the RuvC lid, making similar contacts as the NTS. F999 base stacks with base 23 of the TS upstream of the scissile phosphate. **(j)** Aligned TSps structures with TS in the active site or in the "extended" conformation (gray) to show the RuvC lid transition between cleavage events.

The NTS is positioned upstream between the REC1 helix-loop-helix and RuvC domain, propped up downstream by the NUC stem, and secured across the RuvC active site by a network of positively charged residues (R912, K949, K1072, R1127, R1172, R1226, N1295) (Figure 4.4a and Supplementary Figure C.4e). The narrow pocket between the RuvC and NUC domains forces a bent conformation in the DNA. The NTS engages the RuvC lid α-helix solely via F999 and R1003. F999 base stacks with the NTS upstream of the scissile phosphate, modeled as position 16, and R1003 contacts the bases downstream. The RuvC catalytic residues D908, E993, and D1263 coordinate with the scissile phosphate. We observe two regions of continuous density in the active site that we putatively assign to two Mg^{2+} ions (Figure 4.4b and Supplementary Figure C.4f). This catalytic architecture resembles what has been seen for other class 2 RuvC nucleases with substrate in the active site.[122,169] Similarly, the observed base stacking interaction enforcing a DNA kink adjacent to the scissile phosphate has also been seen in other class 2 RuvC nucleases.[115,122,170–172] Overlaying the 20-bp structure onto the 15-bp intermediate with an occluded active site clearly shows that the RuvC lid regulates active site exposure and must be displaced ∼8 Å to accommodate and engage the NTS in the active site (Figure 4.4c).

To understand the role of the RuvC lid α-helix in catalysis, we tested Cas12a mutants that disrupt lid-DNA contacts seen in the 20-bp structure. We made alanine substitutions of F999 and R1003 and tested cleavage. Surprisingly, there was no defect observed for NTS cleavage of the 20-bp target by the F999A and R1003A mutants (Figure 4.4d and Supplementary Figure C.4g) and we reasoned that a penalty could be masked by other mechanistic features (e.g. network of positive residues holding the NTS across the active

site, rate-limiting R-loop formation). To test this hypothesis, we repeated the cleavage assay with the 16-bp target. NTS cleavage of the 16-bp target showed substantial 25- and 14-fold defects, respectively, for the F999A and R1003A mutants compared with WT Cas12a (0.052 $\pm$0.007 and 0.097 $\pm$0.019 min^{-1}; Supplementary Tables C.3 and C.4), highlighting their role in stabilizing the NTS in the active site, particularly when the NTS is not stabilized by complete R-loop formation.

To observe F999A penalties more directly on RuvC activation, we measured *trans*-cleavage by Cas12a complexed with single stranded activators (Supplementary Figure C.4g). In stark contrast to WT Cas12a, the F999A mutant showed no *trans*-cleavage of the single stranded substrate after 48 h when activated with a 20- or 16-bp R-loop. This complete loss of measurable activity underscores the importance of F999 base stacking to stabilize a DNA substrate within the RuvC active site. Validating structural observations, these cleavage data demonstrate the role of the lid in not only regulating exposure of the RuvC active site, but also promoting DNA binding to the active site.

4.3.7 MD Simulations Probe RuvC Exposure during R-loop Formation

To better understand the role and dynamics of the RuvC lid during initial activation for NTS cleavage, we turned to molecular dynamics (MD) simulations using the adaptively biased MD (ABMD) method.[173,174] This is an advanced enhanced sampling technique that aims at capturing rare events, which we used to investigate the transitioning of the lid from a loop to an α-helix. We took three structures representing different stages of the RuvC lid during R-loop formation: 5-bp R-loop with the lid as an unstructured loop,

the 16-bp R-loop representing a less stable α-helix (forming), and the 20-bp R-loop representing a completed transition into the activated state with an α-helix. Between these structures, we directed the RuvC lid to transition between an unstructured loop and the α-helix as modeled in the 20-bp structure (minima at RMSD ∼1) and determined the associated energetic cost and thus likelihood (Figure 4.4e and Supplementary Figure C.4h).

As expected from our early R-loop structures, the 5-bp free energy profile shows an incredibly high energy barrier of 20 kcal/mol to achieve an α-helical lid. This substantial barrier affirms that the RuvC active site will not spontaneously become exposed without the presence of the DNA substrate. In contrast, the 20-bp structure—as modeled with the α-helical lid engaged with the NTS in the active site—is the most stable state and is reinforced by contacts from the BH, the NTS, and the proximal R-loop. An immediate 6 kcal/mol energetic cost is required to break the helicity of the lid and is associated with the loss of the stabilizing lid contact K1000-E1016 (Supplementary Figure C.4h). Throughout the simulation towards an unstructured loop, the NTS does not leave the active site and remains base stacked to F999.

Consistent with the 16-bp complex representing a more transient state along the R-loop formation pathway, the relevant free energy profile does not reach an α-helical state as stable as the 20-bp complex. Energetic barriers between the minima are smaller (¡1.5 kcal/mol) than those seen for the 20-bp structure. Although the original structure modeled the F999-NTS base stacking interaction, it is lost during the MD simulation and the NTS scissile phosphate moves ∼6 Å away from the active site. Comparison of the 16- and 20-bp MD simulations highlights the importance of the F999 base stacking

interaction to stabilize the NTS in the active site. The complete binding of the NTS in the RuvC pocket also appears to further stabilize the α-helical conformation of the lid, essentially committing the DNA for cleavage. Thus, the expectation would be that DNA product release after RuvC-mediated cleavage would cause the RuvC lid to easily transition back to a more energetically stable unstructured loop and occlude the active site.

4.3.8 Dramatic Conformational Changes Precede TS Cleavage and Reset the Active Site

An outstanding question is how the TS reaches the distant RuvC nuclease to achieve correct polarity and scissile phosphate accessibility.[58,154] MD simulations suggest the REC2 and NUC domains use coordinated movements to guide the TS into the active site55. Single-molecule experiments also point to conformational changes required for TS cleavage, but these changes have not been structurally characterized.[1,42,53,62,154] To provide direct structural insights into this process, we prepared Cas12a with a 20-bp target modified with phosphorothioate (ps) linkages at potential TS cleavage sites to prevent cleavage and product release. From this sample, multiple rounds of three-dimensional variability analyses demonstrated the dramatic domain flexibility involved in RuvC-mediated TS cleavage and captured the TS in the active site (Table 4.1, Supplementary Figures C.1 and C.4).

Most prominently, the post-NTS cleavage states highlight REC2 movements. A substantial class of particles resulted in structures aligned with a fully formed R-loop, cleaved NTS, and no resolved REC2 domain (Supplementary Figure C.5). The NUC

domain swings back and outward up to 12 Å relative to the 20-bp R-loop structure. Density between the NUC domain stem loop and R-loop suggests that the NUC domain remains bound to the R-loop and distal DNA after REC2 dissociation. These structures could represent the recently described Nuc-driven clamped state co-occurring with NTS (and TS) cleavage before DNA release and demonstrate previously hypothesized REC2 dissociation from the R-loop following NTS cleavage.[53]

Next, particles with REC2 docked onto the R-loop were classified according to the relative separation of the REC and NUC lobes into "extended" and "closed" conformations. We present three structures with nominal resolutions ranging from 3.5 to 3.8 Å (Figure 4.4f, and Supplementary Figures C.1 and C.5). When extended, the DNA downstream of the R-loop bypasses the NUC stem loop and bends forward, bringing the TS toward the RuvC active site along the positively charged NUC domain surface. When closed, Cas12a resembles the previous 20-bp R-loop intermediate (RMSD 0.71 Å and 0.57 Å). When aligned by REC1 and the R-loop, the extended structure shows the NUC domain rotating outward by $\sim$11 Å (Figure 4.4g). The shift is also seen with a significant $\sim$7 Å movement of the RuvC domain and minor changes in the REC lobe, demonstrating the breathing motion of the two lobes.

The final reconstruction of the TS in the active site required discarding particles that produced TS heterogeneity caused by the imprecise cleavage behavior that results in a range of cleavage products (Supplementary Figure C.5).[48,58,62,153] Here, R-loop base pairs 18-20 are melted and the single stranded TS is bent into the RuvC active site (Figure 4.4h). The TS is held in place by the same mechanisms used for the NTS: a network of positively charged and RuvC lid contacts via R1003 and F999 (Figure 4.4i). F999 base

stacks with nucleotide 23, consistent with previously published TS cleavage sites.[48,153] F999A TS cleavage of the 20- and 16-bp targets was reduced 9- and 133-fold (0.0857 $\pm$0.0073 and 0.0009 $\pm$0.0002 min^{-1}), relative to that of WT Cas12a (Figure 4.4d and Supplementary Figure C.4g). R1003A had no effect on TS cleavage of the 20-bp target but did reduce TS cleavage of the 16-bp target 7-fold (0.0181 $\pm$0.0045 min^{-1}). The greater defects on TS cleavage underscore the role of the RuvC lid phenylalanine base stacking interaction for securing the TS in the active site.

Interestingly, within the extended structures, the RuvC lid that regulates active site access is no longer an α-helix exposing the active site, but begins to resemble the unstructured loop that occludes the active site in early R-loop intermediates, likely due to the loss of R-loop stabilizing contacts(Figure 4.4j). This observation suggests that the RuvC active site does not remain exposed following initial (NTS) cleavage and resets due to the movement of the two lobes. A RuvC active site that must be reactivated for subsequent cleavage events could explain the greater penalties measured for TS cleavage (*vs.* NTS cleavage) by F999A and R1003A and creates an opportunity to engineer functional nickases.[1,112]

4.4 Discussion

In this work, we have produced a series of novel Cas12a structures capturing DNA interrogation and cleavage that provide a more complete model(Figure 4.5). Cas12a uses a pre-ordered crRNA guide to initiate R-loop formation and relies on the newly docked REC1 domain and a RuvC loop to help stabilize a 5-bp seed. Extension past the

seed requires REC2 eviction and is promoted by newly pre-ordered crRNA nucleotides and REC1-R-loop contacts. Middle R-loop propagation proceeds without any additional contacts from Cas12a to influence the progress of rapid and reversible R-loop formation, likely promoting efficient and stringent target search. Formation of the second helical turn of the R-loop is structurally constrained by the NTS, causing helical distortions that prevent premature REC2 docking and exposure of the adjacent active site.

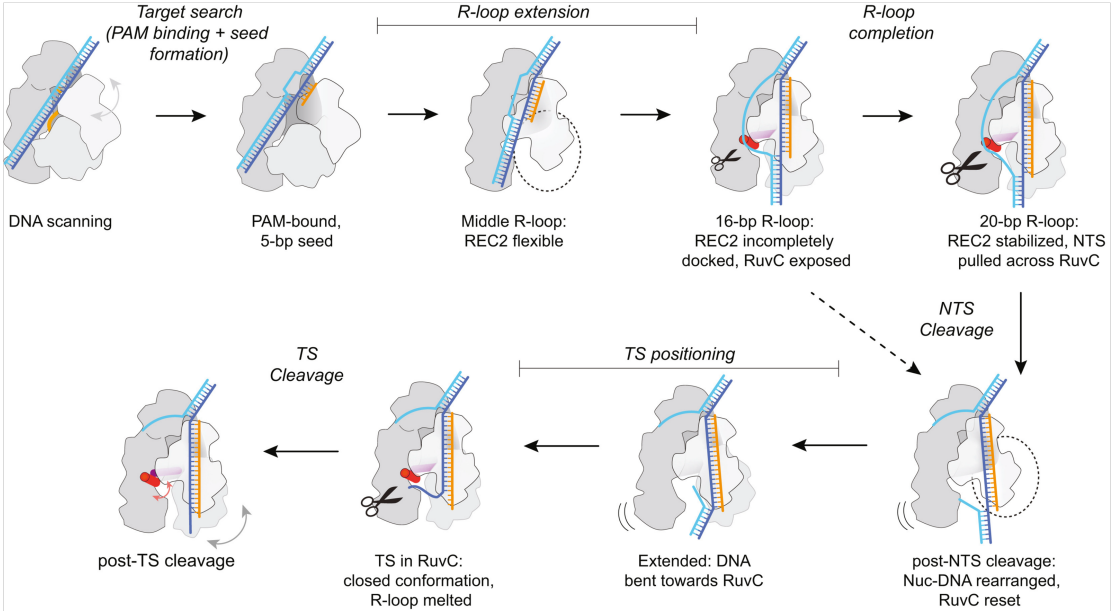


Figure 4.5: **Structural model for Cas12a R-loop formation and DNA cleavage.** Cas12a scans DNA for a PAM sequence with a highly flexible REC1 domain. PAM recognition leads to local DNA melting and seed formation. Cas12a continues to check for guide-target complementarity during R-loop extension and, but limited R-loop contacts ensure an unstable and readily reversible R-loop. At 16 bp, the R-loop is sufficiently long for REC2 to dock and promote RuvC exposure to the NTS via the BH. A complete 20-bp R-loop stabilizes REC2 docking and pulls the NTS across the RuvC active site for efficient cleavage. NTS cleavage leads to REC2 release, providing an opportunity for distal DNA to rearrange with the Nuc domain. As REC2 redocks and the two lobes move closer together, the distal DNA bends toward the RuvC domain, and the RuvC lid exposes the active site to the TS. Following TS cleavage, the flexibility in REC2 and resetting of the RuvC lid are expected to occur repeatedly for *trans*-cleavage. The RuvC lid α-helix uses a critical F999-DNA base stacking to stabilize the scissile phosphate for catalysis.

During formation of the final few R-loop base pairs, the correct heteroduplex shape enables REC2 and BH docking. These three features promote the RuvC lid α-helix transition to expose the active site. In parallel, R-loop completion causes the single stranded NTS to traverse the RuvC domain, displace and engage the lid, and coordinate with the catalytic residues via two Mg^{2+} ions. The RuvC lid that once occluded the active site now plays a role in stabilizing the DNA substrate via a phenylalanine-base stacking interaction. Following NTS cleavage, REC2 becomes highly dynamic, detaching and redocking on the R-loop, the NUC domain becomes flexible too, and the RuvC lid returns to an unstructured loop that occludes the active site. The distal DNA then rearranges with the NUC domain and bends toward the positively-charged pocket of RuvC. As the TS moves towards the active site, REC2 and NUC domains move closer together, returning to the compact, cleavage-competent conformation that stabilizes the TS in the active site. These structural observations are generally consistent with recent MD simulations.[56,113]

Previous structures of Cas12a captured an 8-bp or complete R-loop, identifying R-loop proximity (base pairs 9-11) to the RuvC lid and various REC and BH contacts as important for RuvC activation. Our early and late intermediate structures build on and adjust this model by demonstrating when coordinated changes in Cas12a occur and shedd light on how the nascent R-loop limits activation before sufficient DNA:RNA complementarity is met. The delayed occurrence of stabilizing contacts throughout the R-loop relative to base pair formation establishes R-loop formation and RuvC activation into discrete steps and introduces a conformational checkpoint for Cas12a activation that extends specificity late into the R-loop. These contacts include the “linker” and

"finger" checkpoints previously described for FnCas12a.[1] The dependence on BH-lid contacts, DNA substrate, and the nearby R-loop to stably expose the RuvC active site can explain why previous ternary structures of Cas12a resulted in various RuvC lid conformations.[1,2,16,41,45,72]

Interestingly, our previous analysis of Cas12a off-target activity with single mismatches calculated large energetic penalties on equilibrium binding through most of the R-loop up until position ∼16, within range of penalties expected by nearest neighbor rules for a duplex in solution.[48,175,176] This similarity in behavior would suggest that the R-loop is being formed as it would in solution, with no protein contacts guiding or interfering with the energetics of RNA:DNA base pair formation and DNA:DNA base pair melting. Our new structures that highlight the REC domain flexibility during DNA targeting help rationalize this model for describing R-loop propagation and its inherent reversibility.

The newly-observed flexibility of REC2 and the requirement for REC2 docking to promote efficient DNA cleavage are also cohesive with recent work characterizing PAM-distal mismatches.[177,178] At low Mg^{2+} concentrations, a model in which REC2 docking onto a reversible mismatch R-loop is unfavorable or short-lived and can explain the observed change in rate-limiting step for cleavage of PAM-distal mismatch targets.[48,154] Similarly, a dynamic REC2 paired with reversible R-loop formation and no cleavage to pull the reaction to completion could exhibit a larger observed intermediate R-loop population.

In light of the growing number of related type V nucleases that have been structurally characterized, shared and distinct mechanistic features with Cas12a can be highlighted

to guide rational effector design. Cas12e (CasX) was observed with an unresolved REC2 domain that was characterized to influence DNA cleavge.[179] Cas12m2, a smaller nuclease-dead type V protein, was captured with an intermediate R-loop and an unresolved REC2 domain.[159] These similarities could suggest a shared structural feature of type V R-loop formation that highlights the importance of REC2 flexibility and docking for efficient targeting and nuclease activation.

As SpyCas9 is thoroughly studied and the current gold-standard effector for many gene editing technologies, it is interesting to note shared and distinct features between Cas9 and Cas12a DNA targeting mechanisms. Cas12a appears to deviate in R-loop formation strategies, which likely contributes to its distinct specificity profile. Structures of Cas9 R-loop intermediates suggest that Cas9 contacts R-loop base pairs nearly as they form,[180] in stark contrast to the lack of contacts observed during Cas12a middle R-loop propagation. These structural insights support a model in which Cas9 R-loop formation is less reversible than that of Cas12a and thus less specific against RNA:DNA mismatches during R-loop formation.[48,151–153] Similarly, the appearance of contacts earlier during R-loop propagation could explain Cas9's faster rates of R-loop formation.[90,149] As the R-loop forms the final few base pairs in both Cas12a and Cas9, domain rearrangements and R-loop docking lead to nuclease activation. This mechanistic parallel is a strategy to extend cleavage specificity to late stages of R-loop formation.[81,83,115,168] The work presented here contributes to a more comprehensive mechanistic understanding of Cas12a DNA targeting that rationalizes observed specificity profiles and targeting behavior, draws comparisons to other type V proteins, and highlights underlying mechanistic differences that result in distinct targeting behavior from that of Cas9.

4.5 Acknowledgements

We thank Dr. Axel Brilot and Dr. Evan Schwartz at the Sauer Lab at UT Austin for cryo-EM assistance, Prof. Ken Johnson and Dr. Tyler Dangerfield for use of the Applied Biosystems DNA sequencer, and members of the Taylor Lab for valuable discussion around the project. This work was supported by the National Institutes of Health grants R35GM138348 to D.W.T., R01GM141329 to G.P., and R35 GM131777 to R.R.; National Science Foundation grant CHE-2144823 to G.P.; and Welch Foundation Research Grant F-1938 to D.W.T. Part of this work used Expanse at the San Diego Supercomputing Center through allocation MCB160059 and Bridges2 at the Pittsburgh Supercomputer Center through allocation BIO230007 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Computer time was also provided by National Energy Research Scientific Computing Center (NERSC) under grant no. M3807.

4.6 Materials & Methods

4.6.1 Cas12a Cloning and Purification

Cas12a was cloned into a pET-based expression vector[90] with an N-terminal 6xHis-MBP tag, lac-inducible promoter, and Kanamycin resistance. The His-MBP-Cas12a plasmid was transformed into BL21(DE3) cells (New England Biolabs). A single colony was used to inoculate LB media supplemented with 50 μg/ml Kanamycin for an overnight

culture grown at 37 °C, 200 rpm. The starter culture was then passaged (100X dilution) to 2L of LB supplemented with antibiotic and grown to an OD_{600} of ∼0.6 at which point the culture was chilled and 1 mM IPTG was added to induce expression at 18 °C. Cultures were grown for an additional 20 hours. Cells were pelleted and lysed in equilibration buffer (1 M NaCl, 20 mM HEPES, pH 7.5, 0.5 mM TCEP, 5% glycerol) supplemented with 200 mM PMSF, 0.1% Tween-20, complete Protease inhibitor cocktail tablet (Roche). Lysate was then incubated with 10 mM $MgCl_2$ and 1X DNase I grade II (Roche) at 4 °C with constant shaking for 20 minutes. Lysate was sonicated on ice and clarified by centrifugation at 18k rpm for 30 minutes at 4 °C. Clarified lysate was applied to a pre-equilibrated HisTrap HP column (Cytiva). His-tagged Cas12a was washed with 5% elution buffer (1 M NaCl, 20 mM HEPES, pH 7.5, 5% glycerol, 250 mM imidazole) before elution with a linear gradient of elution buffer. Pooled fractions were digested by recombinant TEV protease (purified in house) to remove the N-terminal His-MBP tag and dialyzed overnight into low salt buffer (150 mM NaCl, 20 mM HEPES, pH 7, 0.5 mM TCEP, 5% glycerol) at 4 °C. Cas12a sample was then run through a HiTrap SP HP column (Cytiva) and eluted by linear gradient high salt elution (1 M NaCl, 20 mM HEPES, pH 7, 0.5 mM TCEP, 10% glycerol). Cas12a was then fractionated over a S200 Increase 10/300 GL column (Cytiva) equilibrated with low salt buffer supplemented with 5 mM $MgCl_2$. After each chromatography step, fractions containing Cas12a were confirmed by SDS-PAGE and pooled. Samples were concentrated to ∼10 μM and aliquots were flash frozen in liquid nitrogen and stored at -80C. Cas12a mutants were cloned using Q5 polymerase and KLD kit (New England Biolabs), sequence verified by Eton Biosciences, and purified the same way as WT AsCas12a. The Loop Deletion

mutant lost residues 1049-1058. The REC2 Deletion mutant lost residues 326-521 with an additional glycine and alanine added.

4.6.2 Target DNA Substrates

DNA oligos were purchased from IDT and resuspended in TE. Target duplexes were formed using 1:1.2 TS:NTS in 50 mM HEPES, 100 mM NaCl and heated for two minutes at 90 °C then slow cooled to 4 °C. When using 5'-FAM-labeled oligos, the unlabeled oligo was added in 1.2-fold excess and labeled duplexes were annealed the same way. DNA targets with reduced complementarity to the guide had PAM-distal mismatches introduced by inverting the TS:NTS base pair so that Watson Crick base pairs in the target DNA were maintained but crRNA:TS base pairs could not form at these locations.

4.6.3 Cryo-EM Sample Preparation, Data Collection

50 μM crRNA (Synthego) was added to an aliquot of purified 12 μM WT Cas12a at a ratio of 1:3 (volume) and incubated at room temperature for 30 min. Equal volumes of 10 μM duplex DNA and assembled Cas12a-crRNA ($\sim$9 μM) were mixed and incubated at 37 ºC or ambient temperature ($\sim$18 °C) before vitrification. DNA binding reaction incubation times varied depending on the DNA substrate used: 8- and 12-bp DNA substrates were incubated for 1 hour at 37 ºC, 16-bp DNA incubated for $\sim$4 minutes at ambient temperature, and 20-bp DNA incubated for 1 minute at ambient temperature. 1.2/1.3R 400 mesh Cu grids were plasma-cleaned for 30 sec in a Solarus 950 plasma cleaner (Gatan). 2.5 μl cryo-EM samples were applied to grids. Using an FEI Vitrobot

MarkIV (Thermo Fisher) set to 4C and 100% humidity, samples were blotted for 6 seconds at 0 force before being plunge frozen into liquid ethane and stored in liquid nitrogen.

The 20-bp dataset was collected on a FEI Glacios cryo-EM microscope (200kV) equipped with a Falcon 4 direct electron detector (Gatan). Movies were recorded in SerialEM[181] with a pixel size of 0.94 Å and a total exposure time of 15 sec resulting in an accumulated dosage and 49e-/Å^2 split into 60 frames. The 8-bp, 12-bp, 16-bp, and TSps datasets were collected on a FEI Titan Krios (300kV) equipped with a K3 Summit direct electron detector (Gatan). Movies were recorded with SerialEM with a pixel size of 0.8332 Å and a total exposure time of 3.8 sec resulting in an accumulated dosage of $\sim$80e-/Å^2 split into 100 frames. Due to preliminary collections resulting in strong orientation bias, datasets were collected at 30° tilt and uploaded to cryoSPARC Live for initial real time processing. For the 8, 12, 16, 20-bp, and TSps datasets, 5,803, 3,492, 12,100, 2,403, and 9,174 movies were collected, respectively. Early attempts to collect a dataset of Cas12a bound to a 4bp target were unsuccessful and not pursued further.

4.6.4 Cryo-EM Data Processing

All datasets were initially processed using MotionCor2[182] and then imported into cryoSPARC for downstream processing, starting with Patch CTF correction[183]. Particles were initially picked according to a minimum and maximum particle diameter of 80 and 150 Å, respectively and manually inspected to limit particle outliers (local power and NCC scores). Final processing workflows are shown in Supplementary Figures C.2 and C.5, and detailed below.

8-bp dataset: 2,748,319 particles were extracted with a box size of 320 pixels with a fourier crop to 128 pixels and then classified into 80 2D classes with a maximum resolution of 8 Å. Classification was done using a class batchsize of 600, 40 online-EM iterations, and two final full iterations, all other settings were kept at default. Manual selection of 2D classes resulted in 1,624,766 particles that went into two rounds of ab-initio volume reconstruction (ab-initio) and hetero refinement (HR). A single class of 719,447 particles was then tested for particle heterogeneity using 3D Variability (3DVar) with resolution filtered to 5 Å and displayed in cluster mode to produce five clusters. The volumes of the two most different clusters were used as input for HR of the 719,447 particles. Nonuniform refinements (NU) with unbinned particles produced structures at 3.35 Å and 3.37 Å (with 246,719 and 224,989 particles, respectively) that still showed particle variability and thus particle stacks from both structures were further classified using 3D classification (3DClass). The 5- and 8-bp 3DClass jobs sorted particles into 10 classes using a O-EM learning rate of 0.5, PCA initialization mode, and 'force hard classification' true. Classes were manually inspected for quality and then refined using NU. The final 5-bp structure reached an overall resolution of 3.42 Å with 60,595 particles. The final 8-bp structure reached an overall resolution of 3.51 Å with 47,461 particles. The 5-bp class without a resolved REC1 domain was further refined to improve resolution using local refinement (LR) with a mask that was generated to excluded REC1 with a 0.05 contour level and dilation radius and soft padding width of 5. The final structure reached an overall resolution of 3.64 Å with 36,668 particles. Local and Global CTF refinements had no effect on map quality.

12-bp dataset: 2,319,655 particles were extracted with a box size of 320 pixels with a fourier crop to 128 pixels and then classified into 100 2D classes with a maximum resolution of 8 Å. Classification was done using a class batchsize of 600, 40 online-EM iterations, and two final full iterations, all other settings were kept at default. Manual selection of 2D classes resulted in 1,509,051 particles that went into two rounds of ab-initio and HR. A single class of 342,296 particles underwent NU with unbinned particles to generate a 3.29 Å consensus structure. 3DVar filtered to 6 Å continued to show heterogeneity in the particle stack so three resulting volumes that represented new 12-bp intermediates with differences in the distal DNA and a 5-bp intermediate were used for HR. Each class was further refined using NU. The final 10-bp structure reached an overall resolution of 3.35 Å with 122,723 particles. Local and Global CTF refinements had no effect on map quality.

16-bp dataset: 5,610,323 particles were extracted with a box size of 320 pixels with a fourier crop to 80 pixels and then classified into 100 2D classes with a maximum resolution of 8 Å. Classification was done using a class batchsize of 400, 40 online-EM iterations, and two final full iterations, all other settings were kept at default. Manual selection of 2D classes resulted in 3,758,114 particles that went into two rounds of ab-initio and HR. A single class of 794,184 particles underwent NU with unbinned particles to generate a 3.12 Å consensus structure. Due to poorly resolved domains and nucleic acid, the particle stack from the consensus refinement was further sorted into 10 classes in a 3DClass job with a 0.6 O-EM learning rate, PCA initialization, 'force hard classification' true, and a solvent mask generated from the earlier NU reconstruction with binned particles. The resulting classes had obvious differences in REC2, R-loop, distal

DNA, and RuvC. Select classes were further refined. The final 15-bp structure reached an overall resolution of 3.27 Å with 158,948 particles. The refined 16-bp structure still showed signs of conformational heterogeneity and therefore contributing particles were further refined using a round of ab-initio and HR to resolve the REC2 domain. The final 16-bp structure reached an overall resolution of 3.22 Å with 97,787 particles.

20-bp dataset: 918,471 particles were extracted with a box size of 320 pixels with a fourier crop to 128 pixels and then classified into 100 2D classes with a maximum resolution of 8 Å. Classification was done using a class batchsize of 600, 40 online-EM iterations, and two final full iterations, all other settings were kept at default. Manual selection of 2D classes resulted in 582,836 particles that went into a single round of ab-initio and HR. A single class with 242,720 particles resembled the expected structure with additional density in the RuvC active site and was refined by NU with unbinned particles to achieve a consensus reconstruction of 3.37 Å. 3DVar showed volumes with and without REC2, so two distinct volumes were used as input for HR sorting. Both classes were further refined by NU. The final 20-bp structure reached an overall resolution of 3.33 Å with 154,794 particles.

TSps dataset: Overall, 5,255,085 particles were extracted from two different collections. Initially, particles were extracted with a box size of 320 pixels with a fourier crop to 128 pixels and then classified into 80 classes with a maximum resolution of 8 Å. From the second collection, particles were extracted with a box size of 320 pixels with a fourier crop to 64 pixels and then classified into 80 classes with a maximum resolution of 8 Å. Particles from selected 2D classes were refined using a single round of ab-initio and HR to remove junk particles. The datasets were merged, and then large

conformational heterogeneity separated out with ab-initio and HR. The structural class lacking the REC2 domain was further refined using HR to demonstrate variability in the nucleic acid components. Of the structural class with a resolved REC2, particles were first separated according the relative separation of the lobes by using 3DVar to observe structural heterogeneity and using the most different maps as templates for HR. Particles were unbinned to produce refined consensus structures of the 'closed' and 'extended' conformations at 3.18 and 3.19 Å, respectively.

3DClass sorted particles of the 'closed' conformation into 10 classes using a O-EM learning rate of 0.5, PCA initialization mode, and 'force hard classification' true. Classes were manually inspected for the density representing TS in the active site and then refined using NU. Due to subtle differences in domain arrangement or RuvC lid configuration, combining classes with exposed RuvC sites did not improve map quality. The 'Exposed RuvC' structure reached an overall resolution of 3.45 Å with 104,326 particles. To improve the quality of density representative of the TS in the RuvC active site, a round of 3DVar with resolution filtered to 5 Å and displayed in cluster mode to produce three clusters. The class with the most prominent defined TS signal was refined via NU. The final TSpsTS structure reached an overall resolution of 3.76 Å with 29,883 particles. 3DVar with resolution filtered at 6 Å sorted particles of the 'extended' conformation into five clusters. Most variability was due to the distal DNA, however classes also differed in extent of separation between the lobes. The most extended conformation was selected as a representative final class and further refined with NU to produce a final structure at an overall resolution of 3.57 Å with 105,267 particles.

4.6.5 Model Building and Refinement

Previously published AsCas12a was rigid body fit into the 20-bp structure within ChimeraX.[184] For structures were rearranged protein, individual domains were rigid body fit. Modeling was performed through iterative rounds of using Isolde[185] and Coot[186] for flexible and local fitting. Models were then subjected to real space refinement within Phenix[187] for final modifications and a validation score.

4.6.6 Cleavage Time Courses

Purified Cas12a and mutants were assembled at 500 nM with cognate crRNA (Synthego) in excess at room temp in assembly buffer (150 mM NaCl, 50 mM HEPES, pH 7, 5 mM $MgCl_2$, 2 mM DTT) for 30 minutes. DNA and Cas12a-crRNA were diluted in reaction buffer which has the same composition as assembly buffer supplemented with 0.2 mg/ml molecular biology grade BSA and pre-warmed to 37 °C. To start the *cis*-cleavage reaction, 5'-FAM-labeled duplexes were combined with Cas12a-crRNA at final concentrations of 50 nM active enzyme and 10 nM duplex DNA and the reaction was carried out in a 37 °C water bath. At various time points, 2 μl were sampled from the reaction and quenched in 4 μl of 0.1 M EDTA. Time points were resolved via capillary electrophoresis using an Applied Biosystems DNA sequencer (ABI 3130xl). Traces corresponding to substrate and product were analyzed to plot fraction cleaved over time. Cleavage time courses were fit using non-linear regression to a single exponential curve (equation: $y = (1-a) * e^{-kx}$; a: amplitude; k: rate) on GraphPad Prism. Rates are reported per minute. Each reaction was repeated at least twice.

Pre-binding competition experiments were performed as described above but with an additional DNA incubation step. 100 nM assembled Cas12a-crRNA was bound to 250 nM unlabeled DNA duplex at 37 °C for the specified amount of time before initiating the cleavage reaction by addition of FAM-labeled PT (20-bp) DNA so that the final concentrations were 50 nM Cas12a-crRNA, 120 nM unlabeled competitor DNA, and 10nM labeled PT DNA.

Activated Cas12a for *trans*-cleavage reactions was assembled as described above. Single stranded activator oligos were incubated with Cas12a-crRNA for 30 minutes to ensure complete binding before addition of radiolabeled *trans* substrate 5'-TTATT initiated the reaction (Supplementary Table C.2). *Trans*-cleavage reactions were initially tested with varying amounts of excess substrate over 10 nM activated WT Cas12a. Replicates were performed at 10 nM Cas12a and 50 nM single stranded *trans* substrate. At various time points, 2 μl were sampled from the reaction and quenched in 4 μl of denaturing quench (60% formamide, 20 mM EDTA, 0.01% bromophenol blue, and 0.01% xylene cyanol). Quenched samples were resolved on a 20% denaturing PAGE gel. Substrate and product intensities were quantitated using a phosphorimager. Cleavage time courses were analyzed using a linear regression to a line (y = k*x+b; k=rate; b=degraded oligo at t=0) in Kaleidograph. Second order rate constants were calculated by dividing the observed cleavage rate by the Cas12a concentration. Additional control experiments (not shown) established that 50 nM substrate was subsaturating. For F999A *trans*-cleavage reactions in which no cleavage was observed, upper limit rates were calculated from an approximate detection limit of 1% cleavage at a 48hr time point. Reactions were repeated with a second *trans* substrate (C_{10}) to validate results.

4.6.7 Plasmid Interference Assay

Cas12a was cloned into a pBAD effector with Ampicillin resistance.[188] The target sequence with an adjacent 5'-TTTA PAM used for in vitro studies was cloned into a modified pACYC vector with Streptomycin resistance and sfGFP expressed from the tac promoter.[165] A crRNA expression cassette was then added to the target plasmid, so that the direct repeat and guide RNA were expressed from the J23119 promoter. Both plasmids were co-transformed (5 ng/plasmid) into a 50 μl aliquot of One Shot BL21(DE3) cells (Invitrogen) using the heat-shock method according to the manufacturer's protocol. Outgrowths at 37 °C were done with addition of 300 μl SOC media for one hour. 150 μl cells were plated onto LB-agar plates containing antibiotics selecting for both plasmids (100 μg/ml Carbenicillin, 100 μg/ml Streptomycin) supplemented with either 1% glucose or 0.2% L-arabinose. Before transformed cells were plated on the arabinose-inducing plates, 2% L-arabinose was added to the liquid culture. Plated cells were incubated at 37 °C for 18-24 hours before colonies were counted. To avoid confounding results from leaky expression of Cas12a, target plasmid interference by Cas12a was compared to a 'no crRNA' control. Colonies were manually counted and calculated colony forming units (CFUs) were plotted. Plasmid interference assays were done in triplicate.

4.6.8 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations were based on three structures of the AsCas12a along the R-loop formation: 5-bp R-loop with the lid as an unstructured loop, the 16-bp R-loop representing a less stable (forming) α-helix, and the 20-bp R-loop representing a

completed transition into the activated state with an α-helix. All systems were embedded in explicit waters and counterions were added to neutralize the total charge, leading to periodic cells of ∼138*149*167 $Å^3$ and ∼307,000 atoms for each system.

MD simulations were performed using the Amber ff19SB force field,[126] with the ff99bsc1 corrections for DNA[127] and the χOL3 corrections for RNA.[101,102] The TIP3P model was employed for water,[104] and the Li & Merz model was used for Mg^{2+} ions.[128] We have extensively employed these force field models in computational studies of CRISPR-Cas systems.[113,129] The Li & Merz model also reported a good description of Mg^{2+} bound sites, in agreement with quantum/classical simulations.[130] An integration time step of 2 fs was employed. All bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm. Temperature control (300 K) was performed via Langevin dynamics,[107] with a collision frequency $\gamma = 1$. Pressure control was accomplished by coupling the system to a Berendsen barostat at a reference pressure of 1 atm and with a relaxation time of 2 ps.[108] The systems were subjected to energy minimization to relax water molecules and counter ions, keeping the protein, the RNA, DNA and Mg^{2+} ions fixed with harmonic position restraints of 300 kcal/mol · $Å^2$. Then, the systems were heated up from 0 to 100 K in a canonical ensemble (NVT), by running two simulations of 5 ps each, imposing position restraints of 100 kcal/mol · $Å^2$ on the above-mentioned elements of the system. The temperature was further increased up to 200 K in ∼100 ps of MD in the isothermal-isobaric ensemble (NPT), reducing the restraint to 25 kcal/mol · $Å^2$. Subsequently, all restraints were released, and the temperature of the systems was raised up to 300 K in a single NPT simulation of 500 ps. After ∼1.1 ns of equilibration, ∼10 ns of NPT runs were carried out allowing the density

of the systems to stabilize around 1.01 g cm^{-3}. Finally, production runs were carried out in the NVT ensemble, collecting ∼120 ns. These simulations were performed using the GPU-empowered version of AMBER 20.[131] The well-equilibrated systems were used as starting points for the subsequent free-energy simulations to study the transitioning of the lid from a loop to the α-helical shape.

4.6.9 Multiple Walker Adaptively Biased MD

The Adaptively Biased Molecular Dynamics (ABMD) method from AMBER[173,174] was used to investigate the transitioning of a loop form of the lid into the α-helical form. This is a non-equilibrium simulation method that belongs to the general category of umbrella sampling methods with a history-dependent biasing potential. Inverse free energy surface (particularly, the potential of mean force; PMF) is built using this potential along with the collective variable that serve as the reaction coordinate. ABMD was implemented in combination with interacting multiple walkers by means of a selection algorithm,[189] replica exchange MD (REMD)[190] and Well-tempered (WT) extension.[191] The WT approach as an enhanced sampling technique is widely used due to its effectiveness and ability to assess convergence of sampling.

In our study, we considered two states of the lid: a loop form of the lid that served as the initial state and the α-helical form of the lid that was considered as the final state. For the 5-bp system, the loop form of the lid was captured in cryo-EM at 3.5 Å resolution and the final α-helical form of the lid was built based on the 20-bp cryo-EM structure. Conversely, the 16- and 20-bp cryo-EM structures, resolved at 3.3 Å and at 3.5 Å resolution respectively, captured the lid in the α-helical form and the loop

form of the lid was built based on the 5-bp state. Here, the root-mean-square deviation (RMSD) of the Cα atoms of the lid (residues 995 to 1009) between the initial and the final state served as the collective variable used to capture the different states of the lid transitioning from an α-helix to a loop. Steered MD was used to obtain eight equidistant states along the CV, which was fed to the aforementioned ABMD method. These simulations were run in a group of eight replicas and the generated weight biases were used to compute the PMF along the collective variable. The simulations were run until convergence (in which the PMF profiles are highly similar with increasing simulation periods (Supplementary Figure C.4h), obtaining 1.2 μs sampling for each system and thereby collecting an overall ensemble of 3.6 μs.

4.6.10 Quantification and Statistical Analysis

Cleavage rates obtained from independent biological replicates were measured and the mean and standard error of the mean of the measurements were calculated and reported. The molecular dynamics simulations were run until convergence (in which the PMF profiles are highly similar with increasing simulation periods).

Chapter V

Enhancement of CRISPR/Cas12a *Trans*-cleavage Activity using Hairpin DNA Reporters

Reprinted with permission from Rossetti M. *et al.*

5.1 Abstract

The RNA programmed non-specific (*trans*) nuclease activity of CRISPR-Cas Type V and VI systems has opened a new era in the field of nucleic acid-based detection. Here, we report on the enhancement of *trans*-cleavage activity of Cas12a enzymes using hairpin DNA sequences as FRET-based reporters. We discover faster rate of *trans*-cleavage activity of Cas12a due to its improved affinity (K_m) for hairpin DNA structures, and provide mechanistic insights of our findings through Molecular Dynamics

simulations. Using hairpin DNA probes we significantly enhance FRET-based signal transduction compared to the widely used linear single stranded DNA reporters. Our signal transduction enables faster detection of clinically relevant double stranded DNA targets with improved sensitivity and specificity either in the presence or in the absence of an upstream pre-amplification step.

5.2 Introduction

The extensive use of CRISPR technology for diagnostic applications has been driven mainly by the discovery of the collateral *trans*-cleavage nuclease activities of CRISPR type V (Cas12) and type VI (Cas13) systems.[59,68,69,192,193] Cas12 and Cas13 proteins are RNA-guided components of bacterial adaptive immune systems and represent an ideal toolbox for molecular diagnostics, as they integrate single- and double-stranded DNA (Cas12) or single-stranded RNA (Cas13) target recognition, signal transduction and amplification in one system.[33,194–196] Their *trans*-cleavage nuclease activity (*trans*-activity) is activated only upon base-pairing of a nucleic acid target to a complementary crRNA guide. In the presence of a FRET-based single-stranded DNA (Cas12) or RNA (Cas13) reporter, consecutive probe degradation events result in an increase of the fluorescence, reflecting the presence of a specific nucleic acid target. So far, this peculiar RNA-programmed *trans*-activity towards single stranded DNA (i.e. Cas12) or RNA (i.e. Cas13) probes has been harnessed by DETECTR and SHERLOCK, respectively, to rapidly detect nucleic acid targets,[17,197–201] and integrated in field-deployable devices for point-of-care use.[17,202–206]

Although CRISPR-based technology has demonstrated vast potential for diagnostic applications, in the absence of a pre-amplification step the direct nucleic acid detection is limited to a detection limit in the picomolar range.[197,207] This restricts the use of amplification-free CRISPR-based assays to the detection of few classes of nucleic acid targets.[208–213] Therefore, most CRISPR-based platforms reported so far generally are combined with an isothermal pre-amplification step (e.g., LAMP, RPA, NASBA, etc.) of the nucleic acid target. However, the addition of this pre-amplification step demands for multiple handling steps and/or high-temperature incubation (55–65 °C), procedures that cannot be easily translated in a Point of Care (PoC) test.[214,215] In addition, the sensitivity of CRISPR-based detection intended as the ratio of the relative change in output to the relative change in input (the steepness of the input/output function) is relatively low, and this poses challenges to the absolute quantification of target also in the range where the cleavage activity directly correlates with target concentration (10-6 M - 10-12 M). Recently, different strategies based on tandem CRISPR nucleases (e.g. Cas13 and Cmr-complex/Csm6) have been successful in achieving amplification-free detection of SARS-CoV-2 RNA.[209,216–218] However, they have been reported only using combination of type III and VI CRISPR–Cas proteins and thus their application is limited to the direct detection of RNA targets. To the best of our knowledge, only one strategy has been reported being able to improve *trans*-activity of Cas12 enzyme until femtomolar limit of detection.[219] It takes advantage of engineered crRNAs showing DNA extension of different length at 3'- or 5'-ends. However, enhancement of *trans*-cleavage activity comes at the cost of increased costs of chimeric DNA/RNA guide. Thus, new strategies

to increase the speed and sensitivity of one-pot detection chemistries of Cas12-based diagnostic systems are required to further optimize its performance.

In this respect, the rational design of alternative DNA reporters with improved affinity for Cas12a proteins, and thus able to enhance the sensitivity and the *trans*-cleavage rate, has been largely neglected. In particular, it is well-recognized that short, single-stranded DNA probes are better performing FRET-based reporters of *trans*-cleavage activity than double stranded DNA.[61,220] Some efforts have been also devoted to the study of different signal transductions,[221,222] but the vast majority of diagnostic assays still rely on the use of linear single-stranded DNA as signal transducers. However, the effect of specific DNA secondary structures (e.g. Hairpin; triplex, etc.) has not been fully investigated.[223–225]

In response to the above considerations, we report here the rational design of single-stranded hairpin DNA reporters bearing a fluorophore and a quencher at the two ends of the self-complementary portions. We first demonstrate that FRET-based hairpin DNA reporters are able to generate faster and enhanced fluorescence signal change compared to linear single-stranded DNA (ssDNA) reporter. Second, we report that Cas12a proteins show higher *trans*-activity on hairpin DNA reporters, and we also provide mechanistic insights on the Cas12a interactions with hairpin substrates. Finally, we demonstrate that hairpin DNA reporters enhance the sensitivity and the specificity of Cas12a-based detection.

5.3 Materials and Methods

Sodium chloride (NaCl), magnesium chloride ($MgCl_2$), tris(hydroxymethyl) aminomethane hydrochloride (Tris-HCl), 2-amino-2-(hydroxymethyl)- 1,3-propanediol (Tris base), boric acid, ethylenediamine tetraacetic acid (EDTA), acrylamide/bis-acrylamide 30% solution, ammonium persulfate (APS), N,N,N',N'-tetramethyl ethylene-diamine (TEMED), were purchased from Sigma-Aldrich, Italy and used without any further purifications. Orange DNA Loading Dye (6x) and O'Range Ruler 5 base pair DNA Ladder were purchased from Thermo Fisher Scientific Italia. EnGen® Lba Cas12a from *Lachnospiraceae* bacterium ND2006 was purchased from New England Biolabs Inc. (Ipswich, UK). EnGen Lba Cas12a is a native protein expressed as a N-terminal 6-His-tagged fusion in *E. coli*. The enzyme has Simian virus 40 (SV40) T antigen nuclear localization sequences (NLSs) at both the N and C-termini of the protein. No mutations are present. Alt-R® A.s. Cas12a (Cpf1) V3 from Recombinant *Acidaminococcus sp.* BV3L6 (A.s.) nuclease, purified from an *E. coli* strain expressing Cas12a was purchased from Integrated DNA Technologies, Inc. (Coralville, UK). Alt-R® A.s. Cas12a (Cpf1) V3 nuclease is high purity, wild-type recombinant *Acidaminococcus sp.* Cas12a. The enzymes include NLSs and C-terminal 6-His tags. No mutations are present. FnCas12a from *Francisella novocida* bacterium was produced as reported in 1.3 Heterologous expression and purification of FnCas12a.

5.3.1 Oligonucleotides

Oligonucleotides (HPLC purified) were purchased from Biosearch Technologies (Risskov, Denmark) and Metabion (Planegg, Germany). All oligonucleotides were dissolved in NE buffer (10 mM Tris-HCl / 50 mM NaCl / 10 mM $MgCl_2$, pH = 7.4) at a concentration of 100 μM and stored at -20 °C. The concentration of the oligonucleotides was determined using Tecan Infinite M200pro (Männedorf, Switzerland) through NanoQuant Plate. Oligonucleotides were annealed before to use by thermocycler. Specifically, DNA oligonucleotides were incubated at 90 °C for 2 min, RNA oligonucleotides at 65 °C for 5 min, both slowly cooled to room temperature (1 °C/min). The sequences of the oligonucleotides used are as follows.

Table 5.1: **FRET-based DNA reporters**

Sequence Name	**Sequence**$^{(a)}$	**Figures**
Linear ss-DNA	5'-(6-FAM)-TTATT–(BHQ1)-3'	5.2, 5.3, 5.4, 5.5, D.1, D.2, D.3, D.4, D.5, D.5, D.7, D.8, D.9, D.10
Stem-loop #5T	5'-(6-FAM)-**CTCTC***ATTTTTA* **GAGAG**–(BHQ1)-3'	5.2, 5.3, D.1, D.2, D.4, D.6, D.8

Continued on next page

continued from previous page

Sequence Name	**Sequence**[(a)]	**Figures**
Stem-loop #10T	5'-(6-FAM)-**CTCTC***ATTTTT* *TTTTTA***GAGAG**-(BHQ1)-3'	5.2, 5.3, 5.4, 5.5, D.1, D.2, D.3, D.4, D.5, D.6, D.7, D.8, D.9, D.10
Stem-loop #30T	5'-(6-FAM)-**CTCTC***ATTTTTTT* *TTTTTTTTTTTTTTTTTTTTA* **GAGAG**-(BHQ1)-3'	5.2, 5.3, D.1, D.2, D.4, D.6, D.8
Variant #1	5'-(6-FAM)-**CTCTC***TTTTT* *TTTTTTT***GAGAG**-(BHQ1)-3'	D.5
Variant #2	5'-(6-FAM)-**CTCTC***ATTATT* *ATTATA***GAGAG**-(BHQ1)-3'	D.5
Variant ss-DNA	5'-(6-FAM)-**TCGTC***ATTTT* *TTTTTTA***CTGAG**-(BHQ1)-3'	D.9

[(a)] Bold bases represent the stem portion, while the italic bases represent the loop portion.

Table 5.2: **Guide crRNA**

Sequence Name	**Sequence**[(a)]	**Figures**
crRNA	5'-UAAUUUCUACUAAGU GUAGAUGUCUACACAUGG CUAAAUCU-3'	5.2, 5.3, D.3, D.4, D.5, D.6, D.7, D.8, D.9
crRNA+ 3'DNA7[(b)]	5'-UAAUUUCUACUAAGU GUAGAUCUCAGGGCGGACU GGGUGCU *TATTATT*-3'	D.3
crRNA-SARS-CoV-2[(c)]	5'-UAAUUUCUACUAAGUG UAGAUAUUGUUACUUUCC UUUACAA-3'	5.5a, 5.5b, D.12, D.13, D.14
crRNA-*Salmonella*[(d)]	5'-UAAUUUCUACUAAGU GUAGAUACCGGCAUCGGC UUCAAUCA-3'	5.5c, 5.5d

[(a)] Underlined bases represent the portion complementary to the ssDNA or dsDNA activator.

[(b)] crRNA+ 3'DNA7 presents a 7-mer DNA extension to crRNA at the 3'-end that has been reported to enhance the sensitivity and specificity of LbCas12a-mediated nucleic acid detection.[219] crRNA+ 3'DNA7 was used in Supplementary Figure D.3.

[(c)] crRNA- SARS-CoV-2 has been designed for the recognition of the S gene SARS-CoV-2.

[(d)] crRNA- *Salmonella* has been designed for the recognition of the *invA* gene.

Table 5.3: **Single- and double-stranded DNA activators**

Sequence Name[(a)]	**Sequence**[(b)]	**crRNAs Used for Targeting**	**Figures**
TS1	5'-AGATTTAGCCATGT GTAGAC*TAAA*-3'	crRNA	5.2, 5.3, D.3, D.4, D.5, D.6, D.7, D.8, D.9
NTS1	5'-*TTTA*GTCTACACAT GGCTAAATCT-3'		
TS- crRNA+3' DNA7	5'-AGCACCCAGTCCG CCCTGAG*TAAA*-3'	crRNA+3'DNA7	D.3
NTS- crRNA+3' DNA7	5'-*TTTA*CTCAGGGCG GACTGGGTGCT-3'		
SARS-CoV-2 ds-DNA amplicon	5'-GATCCTTGTAAAGG AAAGTAACAAT *TAAA*ACCTTC AACACCATTACAAGGTG TGCTAGGATCCCCGGGT ACCGAGCTCGAATTCAC TGGCCGTCGTTTTACA-3'	crRNA-SARS-CoV-2	5.5a, 5.5b

Continued on next page

continued from previous page

Sequence Name[(a)]	**Sequence**[(b)]	**crRNAs Used for Targeting**	**Figures**
	5’- TGTAAAACGACGGCC AGTGAATTCGAGCTCGG TACCCGGGGATCCTAGC ACACCTTGTAATGGTGT TGAAGGT*TTTA*ATTGTTA CTTTCCTTTACAAGGAT C–3’		
PM	5’-TTGTAAAGGAAAGT AACAAT*TAAA*-3’	crRNA-SARS-CoV-2	5.5e, 5.5f, D.12, D.13, D.14
MM1	5’-TTGTAAAGGAAAGT AACAA**C***TAAA*-3’		
MM4	5’-TTGTAAAGGAAAGT AA**T**AAT*TAAA*-3’		
MM8	5’-TTGTAAAGGAAA**T**T AACAAT*TAAA*-3’		
MM12	5’-TTGTAAAG**T**AAAGT AACAATTAAA-3’		

Continued on next page

continued from previous page

Sequence Name[a]	**Sequence**[b]	**crRNAs Used for Targeting**	**Figures**
MM16	5’-TTGT**G**AAGGAAAGT AACAATTAAA-3’		
MM20	5’-**G**TGTAAAGGAAAGT AACAAT*TAAA*-3’		
MM1-2	5’-TTGTAAAGGAAAGT AACA**TC***TAAA*-3’		
MM4-5	5’-TTGTAAAGGAAAGT A**GT**AAT*TAAA*-3’		
MM8-9	5’-TTGTAAAGGAA**GT** TAACAAT*TAAA*-3’		
MM12-13	5’-TTGTAAA**CT**AAAGT AACAAT*TAAA*-3’		
MM16-17	5’-TTG**GG**AAGGAAAGT AACAAT*TAAA*-3’		
MM19-20	5’-**GA**GTAAAGGAAAGT AACAAT*TAAA*-3’		
NTS	5’-*TTTA*ATTGTTACTTTC CTTTACAA-3’		

Continued on next page

continued from previous page

Sequence Name[a]	**Sequence**[b]	**crRNAs Used for Targeting**	**Figures**
Salmonella ds-DNA amplicon	5'-TGGCGATATTGGTGT TTATGGGGTCGTTCTA CATTGACAGAATCCTC AGTTTTTCAACGTTTC CTGCGGTACTGTTAAT TACCACGCTCTTTCGT CTGGCATTATCGATCA GTACCAGCCGTCTTAT CTTGATTGAAGCCGAT GCCGGTGAAATTATCG CCACGTTCGGGCAATT CGTTATTGGCGATAGC CTGGCGGTGGGTTTTG TTGTCTTCTCTATTGT CACCGTGGTCCAGTTT ATCGTTATTAC-3'	crRNA-Salmonella	5.5c, 5.5d

[a] TS represent the target single-stranded DNA, complementary to the RNA guide, whereas NTS represent the complementary non-target single-stranded DNA. PM stands for perfect match and MM for mismatch (reported in bold in the sequence). The numbers

following “MM” indicate the position of the mismatches respect to the PAM motif (i.e. MM1 indicates that mismatch is the first base following PAM motif, MM4 indicates that mismatch is the fourth base following PAM motif, etc).

$^{(b)}$ Italic bases represent the PAM motif; bold bases represent the mismatches.

5.3.2 Heterologous Expression and Purification of FnCas12a

The protocol provides by Mohanraju *et al.* was followed.[145] The pDS015 plasmid, containing the *Francisella novocida* Cas12a gene was used for the heterologous expression of the protein in *E. coli* BL21(DE3) Rosetta™ 2 (Merck, Novagen), as fully described in the relative Bio-Protocol site (www.bio-protocol.org/e2842). Starting from 16 g of induced wet cells, we obtained ca. 10 mg of 95% purified protein, as confirmed by SDS-PAGE analysis. Aliquots of FnCas12a were finally stored at -20 °C.

5.3.3 DNA Preparation of Plasmid Containing the SARS-CoV-2 Sequence (S gene)

The pUC19 plasmid was chosen as recipient vector to clone the sequence of the S gene of SARS-CoV-2.[226] The entire sequence was obtained from Eurofins Genomic (Germany) as oligonucleotides pairs flanked by a short sequence (underlined) which simulate, after oligonucleotide annealing, protruding ends after the *BamH* I digestion, for the subsequent cloning procedures.

NTS-Fwd: 5’-GATCCTAGCACACCTTGTAATGGTGTTGAAGGTTTTAATTGTTA

CTTT

CCTTTACAAG-3’

NTS-Rev: 5'-GATCCTTGTAAAGGAAAGTAACAATTAAAACCTTCAACACCATT

ACAA

GGTGT GCTAG-3'

The pUC19 plasmid was also digested with *BamH* I enzyme, and the resulting DNA fragment was gel-purified by using NucleoSpin® Gel and PCR Clean-up® kit from Macherey-Nagel GmbH (Germany). The annealed oligonucleotides were mixed with the digested plasmid and incubated in the ligation mixture with a plasmid:oligos 1:10 ratio. Afterwards, the mixture was used to transform *E. coli* DH5a competent cells and positive white colonies were selected on Ampicillin selective Luria-Bertani (LB) Agar plates supplemented with 5-bromo-4-chloro-3-indolyl-b-D-glucopyranoside (X-Glc) and Isopropyl β-d-1-thiogalactopyranoside (IPTG). One positive colony was used for the plasmid preparation, achieving the final pUC19-BZ construct.

5.3.4 Production of SARS-CoV-2 Amplicons through PCR

In order to produce different amount of target SARS-CoV-2 amplicons for the CRISPR-based detection assay, six aliquots containing known concentrations of pUC19-BZ variant plasmid (from 20 nM to 20 aM) were amplified by PCR with M13-Fwd (5'-TGTAAAACGACGGCCAGT-3') / NTS-Rev oligonucleotides, under the following conditions: an initial denaturation at 95 °C for 5 min, 30 cycles of 30 s at 95 °C, 30 s at 50 °C and 30 s at 72 °C, and a final extension of 5 min at 72 °C. All amplified DNA products were cleaned by the NucleoSpin® Gel and PCR Clean-up® kit from Macherey-Nagel GmbH (Germany), in order to remove oligonucleotides, free nucleotides and the DNA polymerase enzyme.

5.3.5 Preparation and Standardization of *Salmonella* Strain

The sensitivity test was carried out using *Salmonella typhimurium* ATCC 14028 (Oxoid Ltd, Basingstoke, UK). The strain was revitalized by incubation in 10 ml of Tryptone Soy Broth medium (Lab M Ltd, Heywood, UK) for 24 h at 37 °C. Furthermore, the strain was standardized by determining the bacterial load with the method of serial dilutions and sowing by inclusion on Plate Count Agar (Lab M Ltd, Heywood, UK) that yielded about 108 CFU/ml. An aliquot (2 ml) of strain was centrifuged at 14,000 rpm for 10 minutes, the pellet was suspended in 1 ml of distilled water and centrifuged at 14,000 rpm for 10 minutes, and the pellet was suspended in 200 μl of 6% Chelex. The suspension was incubated for 8-10 minutes at 100 °C, finally placed on ice for 1 minute. The suspension was then centrifuged (5 minutes at 14,000 rpm) to allow the Chelex, cell components, membrane residues and cell wall to settle to the bottom of the tube. 100 μl of the supernatant was carefully transferred to a new tube and used as a template in molecular methods.

5.3.6 PCR Amplification of *Salmonella* Samples

Briefly, the primer used amplify a highly specific segment of the region of *invA* gene, conserved in all salmonella serotypes. The reaction mixture consists of a total of 50 μl divided as follows: 25 μl of master mix (QuantiTech Multiplex PCR Master Mix NoROX, QIAGEN), 1 μl (10 μM) of forward primer F3 (GGCGATATTGGTGTTTATGGGG), 1 μl (10 μM) of reverse primer B3 (5'- AACGATAAACTGGACCACGG-3', Hara-Kudo *et al.*, 2005), 5 μl of molecular biology water, 4 μl of DNA template. The reaction was run at 95 °C for 15 min, followed by 40-cycle at 95 °C for 30 s and 60 °C for 30 s.

All amplifications were performed using the GeneAmp PCR System 9700 instrument (Applied Biosystems). PCR products were analysed by 2% agarose gel electrophoresis using 5 μl of ethidium bromide (10 ng/ml) and run at 90 V for 60 min and finally visualized by UV transillumination. The sizes of the sample PCR fragment (244 bp) were determined comparing their position on the gel with those belonging to the standard DNA ladder (EZ Load 100 bp molecular ruler, Bio-Rad Life Science Research). Finally, the PCR products were purified with kit QIAquick® PCR Purification KIT (QIAGEN).

5.3.7 Melting Assays and Thermodynamic Analysis

Fluorescence melting curves (Figure 5.2a and Supplementary Figure D.1) were obtained by heating a solution containing 50 nM of DNA reporter in NE buffer (10 nM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, pH = 7.9) from 20 °C to 90 °C at a rate of 1 °C min^{-1} and monitoring fluorescence at every 0.5 °C, setting $\lambda_{ex.}$ = 488 nm and $\lambda_{em.}$ = 520 nm. All the solutions were annealed prior to analysis by heating the solutions to 90 °C and slowly cooling down at room temperature. The reported melting curves have been normalized through the use of the interpolation model that allows to estimate the melting temperature (T_m) for each experiment (Figure 5.2a and Supplementary Figure D.1).[227] To estimate the switching equilibrium constant (K_s) of the hairpin DNA reporters, we assumed a two-state model and ΔH° independent of temperature ($\Delta C_p = 0$).[227,228] Firstly, we calculated the folded and unfolded fraction of the hairpin reporters as function of temperature according to the following equation:

$$\alpha = (B_U - F_T)/(B_U - B_F) \tag{5.1}$$

where α represents the folded fraction, B_U and B_F are the baseline fluorescence values of the unfolded and folded species, respectively and, F_T is the fluorescence signal of the hairpin DNA reporter at a given temperature. K_s was then calculated as $K_s = (1-\alpha)/(\alpha)$, by taking in accounts the α between 0.15 and 0.85. Van't Hoff analysis was carried out by plotting $ln(K_s)$ as a function of $1/T$ (Supplementary Figure D.1) and data were fitted by using the following equation:

$$ln(K_s) = \frac{-\Delta H^\circ}{R} * \frac{1}{T} + \frac{\Delta S^\circ}{R} \tag{5.2}$$

We assumed that all the variants by sharing the same stem portion (related to the enthalpy contribution) present the same average ΔH° value (44.1 ± 2.8 kcal/mol). Finally, K_s, was calculated as $K_s = \exp(-\Delta G^\circ/RT)$ at 37 °C.

5.3.8 Fluorescence *Trans*-cleavage Assays

Double-stranded DNA activator was prepared by annealing 3-fold molar excess of the NTS to TS in buffer (10 nM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, pH = 7.9) heating at 90 °C for 2 min, and slow-cooling on the benchtop. *Trans*-cleavage assays were performed as follow, unless otherwise stated. 4.5 μl of Cas12a/cRNA complex (200 nM) previously incubated with the single (ssDNA) or double stranded (dsDNA) DNA activator (10 nM) for 30 min at 37 °C, are diluted into a final volume of 45 μl containing 100 nM DNA-reporter (final concentration) at 37 °C. We performed all steady state fluorescence measurements using a Cary Eclipse Fluorimeter. We recorded emission spectra with excitation at 488 nm and acquisition from 500 nm to 600 nm. We fixed the

excitation and emission bandwidths to 5 nm in all the experiments. Kinetics experiments were performed by fixing excitation at λ_{exc} = 488 nm and acquisition at λ_{em} = 520 nm, and setting the excitation and emission bandwidths to 5 nm in all the experiments. Background-corrected fluorescence values were calculated by subtracting fluorescence values obtained from reactions carried out in the absence of activator at the specific time of analysis.

To determine the cleaved fraction of substrate over time (Figure 5.3b and Supplementary Figure D.6), fluorescence kinetics were fitted to a single exponential decay curve (Kaleidagraph, Software), according to the following equation:

$$Fraction\ cleaved = A * (1 - \exp(-kt)) \tag{5.3}$$

where A is the amplitude of the curve, k is the first-order rate constant, and t is time. Calibration curves performed for sensitivity assays (Figure 5.5a) have been obtained after 1 h of incubation of different concentrations of ds-DNA amplicons in a solution containing LbCas12a/crRNA (20 nM, final concentration) and the Linear-ssDNA or the Stem-loop #10T DNA reporter (100 nM, final concentration), as previously reported. The detection of PCR-amplified amplicons (Figure 5.5b) have been performed by diluting the PCR products 500 times in NE buffer and then by transferring 5 μl to a vial containing 25 μl of 200 nM LbCas12a/crRNA-SARS-CoV-2. The resulting solutions were incubated for 30 min at 37 °C. After that, 45 μl of DNA reporter (final concentration 100 nM) was added and incubated for 1 h at 37 °C. Finally, the solutions were heated at 65 °C for 10 min and cooled down to 37 °C, in order to deactivate LbCas12a

and *trans*-cleavage reaction. Detection of Salmonella (Figure 5.5c,d) was performed by adding 2.5 μl of 200 nM LbCas12a/crRNA (RNP) complex to the 25 μl of PCR products at different dilution factors in NE buffer containing a 500 nM final concentration of DNA reporter and kinetic were followed for 1 h at 37 °C, by using a Tecan M200pro plate reader using top reading mode with black, flat bottom non-binding 96-well plates. Kinetic experiments performed for testing the specificity, were carried out by adding 4.5 μl of activator (final concentration 10 nM of TS or MM + 30 nM of NTS) in 45 μl of solution containing 100 nM of reporter and 20 nM of LbCas12a/ crRNA-SARS-CoV-2 (previously incubated for 30 min at 37 °C at 200 nM concentration) (Figure 5.5e,f and Supplementary Figure D.12).

5.3.9 Michaelis-Menten Enzyme Kinetic Measurements and Data Analysis

Trans-cleavage reactions were measured by adding 4.5 μl of a 10x concentrated solution of active complex (10 nM of ds-DNA activator + 200 nM of LbCas12a/crRNA complex pre-incubated for 30 min at 37 °C) to a solution (final volume of 45 μl) of FRET-based DNA reporters at different final concentrations (from 23 nM to 1.9 μM). Specifically, the final concentrations of LbCas12a/crRNA and ds-DNA activator are 20 and 1 nM, respectively, so that the concentration of the effective RNP complex is 1 nM. Reactions were carried out in three replicates and fluorescence measurements were acquired every 8 ms (λ_{ex}: 488 nm; λ_{em}: 520 nm). We calculated the initial reaction velocities (V_0) for each concentration of reporters by fitting to a linear regression the data collected during the first 180 s. We converted the calculated reaction velocities from a.u./min to nM/min

by assuming that fluorescence $F(t)$ measured during *trans*-cleavage experiments is the sum of the fluorescence from cleaved and uncleaved reporters,[207] as reported in the following equation:

$$F(t) = F_{cleaved}(t) + F_{uncleaved}(t) \tag{5.4}$$

where $F_{cleaved}(t))$ and $F_{uncleaved}(t)$ are the fluorescence from cleaved and uncleaved reporters, respectively. Fluorescence signal of fully cleaved reporters ($F_{cleaved}$) at the reported concentrations were obtained by incubating 45 μl of reporter for 24 h at 37 °C in the same experimental conditions reported above by assuming the *trans*-cleavage reaction completed. Fluorescence signal of uncleaved reporters ($F_{uncleaved}$) was obtained under the experimental procedure by using an inactive LbCas12a/crRNA complex (in the absence of the dsDNA target). We obtained calibrations curves of $F_{cleaved}$ and $F_{uncleaved}$ versus increasing concentrations of reporters (from 24 nM to 1.9 μM) by fitting data to a linear regression (see Supplementary Figure D.8).

Since we know that at time (t) the overall fluorescence is the sum of $F_{cleaved}(t)$ and $F_{uncleaved}(t)$ which is function of the reporter concentrations ($[C]$), we can assume that:

$$F(t) = m1 * [C_{cleaved}] + m2 * (1 - [C_{cleaved}]) \tag{5.5}$$

where $m1$ and $m2$ represent the calibration curve equations (i.e. angular coefficient) from Supplementary Figure D.8c. The reaction velocities in nM/s is obtained by differentiating the above reported equation with respect to time. The calculated reaction velocities (nM/min) were plotted versus FRET-based DNA reporter (substrate) concentration and fitted to the Michaelis-Menten equation using GraphPad Prism software

(GraphPad, CA, USA) to obtain V_{max} and K_m, according to the following equation: $Y = (V_{max} * X)/(K_m + X)$, where X is the substrate concentration and Y is the enzyme velocity. Finally, turnover number (k_{cat}) was calculated by using the following equation: $k_{cat} = V_{max}/E_t$, where E_t is the concentration of the active ribonucleoprotein complex = 1 nM.

5.3.10 Native Polyacrylamide Gel Electrophoresis (PAGE) Assays

Native PAGE experiments were obtained with 18% polyacrylamide (29:1 acrylamide/bisacrylamide) in TBE 10x buffer (1 M Tris, 0.9 M boric acid, and 0.01 M EDTA), pH 8.3. Gel solution were prepared by mixing 5.7 ml of diH_2O, 4.1 ml of TBE 10x buffer, 4.2 ml of 30% acrylamide/bisacrylamide solution, 75 μl of 10% APS, 14 μl of TEMED. A volume of 10 μl of each sample was mixed with 1 μl of 60% (w/w) Glycerol and then the mixture was added into the gel for the electrophoresis assay. O'Range Ruler 5 base pair DNA Ladder was used as the DNA standard. The samples were obtained following the same procedure reported for fluorescence *trans*-cleavage assay using 250 nM of FRET-based DNA reporter. The reactions were stopped at different times (0, 15, 30, 45, 60, 120 and 180 min), by heating the solutions for 10 min at 65 °C in order to inactivate the Cas12a enzyme. The native PAGE was carried out in a Mini-PROTEAN Tetra cell electrophoresis unit (Bio-Rad) at room temperature, at a constant voltage of 120 V, using TBE 1x buffer (0.1 M Tris, 0.09 M boric acid, and 0.001 M EDTA) at pH 8.3 for 80 min.

5.3.11 PAGE Analysis of *Trans*-cleavage Digestion of FRET-based DNA Reporters

Gel was scanned by a ChemiDoc Imaging System (Bio-Rad) that allows to image the fluorescent bands (FAM emission) associated to the cleaved and uncleaved FRET-based DNA reporters. However, the same gels were imaged also after 30 minutes of staining in SYBR gold (Invitrogen) dissolved in a TBE 1x buffer at pH 8.3 to visualize the ladder signal. We calculated the cleavage fraction (α) according to the following equation:

$$\alpha = \frac{F_t - F_{uncleaved}}{F_{cleaved} - F_{uncleaved}} \tag{5.6}$$

where $F_{uncleaved}$ and $F_{cleaved}$ are the intensity of the bands associated to the reporter at reaction time equal to 0 s and 180 minutes, respectively, and $F(t)$ is the fluorescence intensity associated to the cleaved reporter at a given time of analysis.

5.3.12 Structural Models

Molecular simulations were based on the X-ray structure of the CRISPR-Cas12a system in complex with RNA and DNA (PDB ID: 5NFV),[2] as obtained from Francisella novicida at 2.5 Å resolution. The linear DNA was modelled from the DNA target strand of this X-ray structure, by mutating the sequence to 5'-CTCTCATTTTTAGAGAG-3', following the experimental studies. The hairpin DNA was modelled from an NMR solved hairpin DNA (PDB ID: 1QE7),[229] by mutating the sequence same as the Linear-ssDNA mentioned above. The three systems were embedded in explicit waters, while Na^+ ions were added to neutralize the total charge, leading to an orthorhombic periodic

cell. The simulation cell for the CRISPR-Cas12a complex was of $\sim 120 \cdot 130 \cdot 150$ Å containing a total of ∼210,000 atoms. The simulation cell of both the linear and stem loop DNA had a dimension of $\sim 90 \cdot 60 \cdot 60$ Å^3, containing ∼25,000 atoms.

5.3.13 Molecular Simulations

All-atom molecular dynamics (MD) simulations were performed using the Amber ff12SB force field, which includes the ff99bsc0[100] corrections for DNA and the ff99bsc0+χOL3[101,102] corrections for RNA. The Allnér force field[103] was used for Mg^{2+} ions and the TIP3P model was used for water. These force field parameters and the simulation protocol are well-established for protein/nucleic acid complexes, as also shown in our recent studies of CRISPR-Cas9,[73–75,88,94] corroborated by NMR experiments[94] and quantum mechanical calculations.[78] An integration time step of 2 fs was employed and all bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm. Temperature control was performed at 300 K via Langevin dynamics[107] with a collision frequency of γ = 1/ps. Pressure control was accomplished by coupling the system to a Berendsen barostat[108], at a reference pressure of 1 atm and with a relaxation time of 2 ps. The systems were subjected to energy minimization to relax water molecules and counter ions, keeping the protein, RNA, DNA, and Mg^{2+} ions fixed with harmonic position restraints of 300 kcal/mol·Å^2. The systems were then heated from 0 to 100 K in the canonical ensemble (NVT), by running two simulations of 5 ps each while imposing position restraints of 100 kcal/mol·Å^2. The temperature was further increased up to 200 K in a ∼100 ps MD run in the isothermal–isobaric ensemble (NPT), reducing the restraint to 25 kcal/mol·Å^2. Subsequently, all restraints

were released, and the temperature of the system was raised to 300 K in a single NPT simulation of 500 ps. After ∼1.1 ns of equilibration, a ∼10 ns of NPT run was carried out for the density of the system to stabilize at around 1.01 g/cm^3. Production runs for the CRISPR-Cas12a system were carried out reaching ∼4 μs of MD simulations in two replicates, gaining ∼8 μs of collective ensemble. After these simulations, two structures, displaying an open conformation of the system, were extracted for further ∼200 ns of conventional MD to obtain an ensemble of conformations for molecular docking. The step loop DNA and linear DNA were also simulated for ∼200 ns of MD in two replicates, starting from different configurations and velocities. Molecular simulations have been carried out using the GPU-accelerated version of AMBER 19.[109] The most representative configurations were selected from the top three clusters, as obtained by performing k-means clustering on the collected ensemble of the CRISPR-Cas12a complex, stem loop DNA and linear DNA. These selected configurations of the DNA models and of the CRISPR-Cas12a system were object of docking calculations as described below.

5.3.14 Docking and Binding Energies

We performed rigid body docking of the stem loop and linear DNA models as ligands for the target CRISPR-Cas12a complex using HADDOCK2.4.[230] The catalytic core composed of the DED motif (D916, E1005, D1254) in Cas12a were treated as the "active" residues for HADDOCK2.4 docking, while the solvent-exposed residues within 6.5 Å of the active residues were treated as "passive" residues. The five dT-repeats in the DNA models were treated as active residues and passive residues were selected following the criterion mentioned above. Both the DNA models were restrained to

preserve their structure while docking. For the all-atom representation, OPLSX non-bonded parameters are used both for the protein[133] and DNA.[133] The most representative docked conformations from the top cluster, based on the electrostatic and Van der Waal energies, and HADDOCK2.4 score, were selected for further analyses. The docked structures were minimized and equilibrated as reported above (Molecular Simulations paragraph) for subsequent ∼200 ns of conventional MD simulations. We calculated the binding free energy for the association of DNA with Cas12a using the Molecular Mechanics with Generalized Born and Surface Area (MM-GBSA) method.[229] Binding free energy calculations where performed over the ∼200 ns ensemble, in windows of ∼20 ns. The GBn model described by Mongan, Simmerling, McCammon, Case and Onufriev was employed.[229]

5.4 Results and Discussion

5.4.1 Design of FRET-based Hairpin DNA Reporters

To enhance the *trans*-cleavage activity and the signal transduction of Cas12a, we have investigated the use of hairpin DNA molecules as fluorescent reporters. Consisting of a stem-loop DNA modified at the two self-complementary ends with a fluorophore/quencher pair, they are a class of synthetic biomolecular switches widely used for live-cell analysis of native RNA transcripts,[231,232] *in-vitro* diagnostics[233–236] and intracellular analysis at the single-molecule level,[237] and thus represent an ideal test bed for our studies. The simple hairpin structure indeed separates the single-stranded portion that preferentially can be cleaved by activated Cas12a proteins (the loop), from

those that define the switching equilibrium and signal transduction (the stem). From the perspective of signal transduction, one crucial advantage of using stem-loop DNA reporters over conventional short ssDNA probes is the superior quenching efficiency and the relatively low fluorescence background due to the tight contact between the fluorophore and the quencher.[238–240] In addition, we hypothesized that if Cas12a is able to cleave the single stranded loop portion, the short self-complementary stem should not be stable enough to maintain the non-emissive hairpin conformation, thus leading to a rapid increase of the fluorescence signal (Figure 5.1).

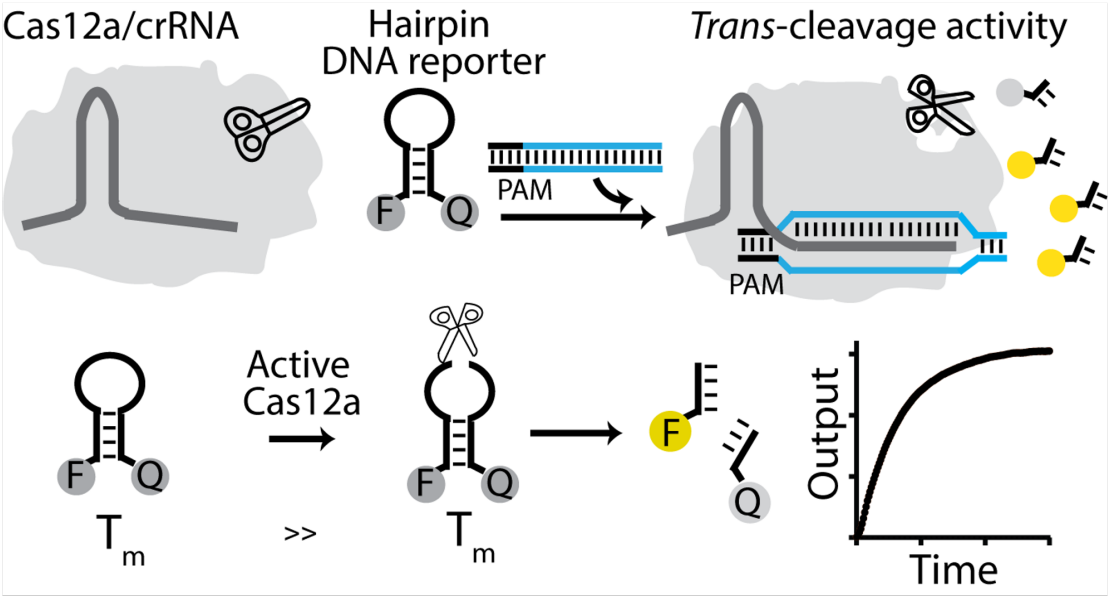


Figure 5.1: **Schematic for Cas12a *trans*-cleavage activity based on hairpin DNA reporters.** We explore the use of different FRET-based hairpin DNA reporters as signal transducers of the Cas12a-based *trans*-cleavage reaction. The hairpin structure separates the single-stranded portion that preferentially can be cleaved by activated Cas12a proteins (the loop) from those that define the switching equilibrium and signal transduction (the stem). Once the ternary complex of Cas12a/crRNA/dsDNA is formed, the *trans*-cleavage activity is activated and the hairpin DNA reporter can be cleaved, thus generating a fluorescence signal change.

To demonstrate our hypothesis, we have designed a set of three hairpin DNA reporters that retain a common stem sequence (i.e. 3GC+2AT) and fluorophore/quencher pair

(6-FAM/BHQ-1), but present loop domains of different length (Figure 5.2a, left). In this way, it is possible to finely modulate the switching constant (K_s) associated to the equilibrium between the non-emitting hairpin configuration (OFF-state) and the emitting unfolded state (ON-state) of the DNA reporter.[241,242] Specifically, by increasing the loop length (composed by thymine repetitions) from 5T (Stem-loop #5T), to 10T (Stem-loop #10T) and 30T (Stem-loop #30T) we can rationally decrease the thermal stability of the hairpin DNA reporters (Figure 5.2a, right). Indeed, melting assays confirm that the melting temperature at which they unfold strongly correlates to the loop length ($T_{M,\,Stem-loop\,\#5T}$ = 72.4 $\pm$ 0.3 °C; $T_{M,\,Stem-loop\,\#10T}$ = 67 $\pm$ 1 °C; $T_{M,\,Stem-loop\,\#30T}$ = 53.3 $\pm$ 0.5 °C, Figure 5.2a, right). From the thermodynamic analysis of the data we confirmed that a longer loop length corresponds to a higher switching constant ($K_{s,\,\#30T}$ = (2.9 $\pm$ 0.2) x 10^{-2}), resulting in a destabilized non-emitting folded state and a lower signal gain due to the increased fluorescence background (Supplementary Figures D.1 and D.2). In contrast, a shorter loop produces an overly stabilized structure ($K_{s,\,\#5T}$ = (7 $\pm$ 2) x 10^{-4}) and a higher signal change. As expected, the linear single-stranded FRET-based reporter used in DETECTR (Linear-ssDNA) does not show any fluorescence signal increase along with temperature, in agreement with the absence of secondary structures. Of note, we fixed the length of the stem because longer stems might lead to the formation of hairpin-hairpin dimers because of the self-complementary nature of the 5'- and 3'- ends.

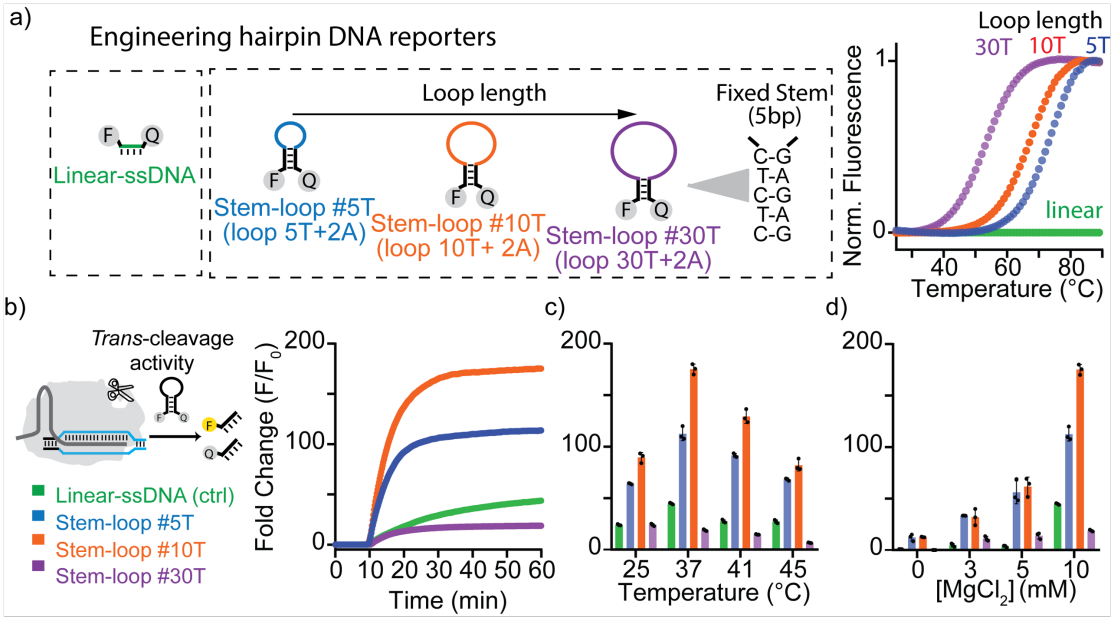


Figure 5.2: **Design and characterization of different hairpin DNA reporters as FRET-based reporters of LbCas12a *trans*-cleavage activity.** **(a)** We tested three different hairpin DNA reporters having same stem composition (3GC+2AT) but different loop length and an increasing number of thymine (from 5 to 30). All the DNA reporters were functionalized with 6-FAM at 5'-end and with BHQ-1 at 3'-end. Melting curves indicate that by increasing the number of thymine in the loop we can rationally control the melting temperatures of the reporters and thus the stability of the hairpin structure. **(b)** LbCas12a *trans* nuclease activity on linear ssDNA and hairpin DNA reporters. The fluorescence kinetic analysis over time of LbCas12a *trans*-cleavage activity triggered by a fully complementary 20bp-long double stranded DNA activator show enhanced signal transduction using the Stem-loop #10T reporter. **(c)** Fluorescence fold change as a function of the temperature after 1h of LbCas12a *trans*-cleavage activity. **(d)** Effect of Mg^{2+} concentration on LbCas12a-based *trans*-cleavage activity. The assays were performed by adding 4.5 μl of LbCas12a/crRNA complex (200 nM) preassembled with the dsDNA activator (10 nM) to a solution containing the DNA reporter (100 nM). The *trans*-cleavage reactions in c) and d) were stopped after 1 h by heating the solutions at 65 °C for 10 min. All bar graph shows the mean fold change calculated as $F\text{-}F_0/F_0$ (background-subtracted fluorescence divided by background fluorescence of the DNA reporter at time t = 0) $\pm$ SD, where n = 3 replicates.

5.4.2 Faster and enhanced Signal Change Using Hairpin DNA Reporters

As our test bed for the study of signal transduction, we measured the a-specific nuclease activity of *Lachnospiraceae bacterium* Cas12a (LbCas12a) using the three FRET-based hairpin reporters and the Linear-ssDNA. Consistent with our hypothesis, all DNA reporters display increased fluorescence upon *trans*-activity thus confirming that LbCas12a protein can degrade also DNA reporters folded in a hairpin structure. Notably, the Stem-loop #10T yields the highest fluorescence signal ($\sim$200-fold change) compared to the other DNA reporters (Figure 5.2b). The lower fold signal change obtained using Stem-loop #30T reporter can be ascribed to the higher fluorescence background at 37 °C, in agreement with the T_M of melting of this reporter (Figure 5.2a and Supplementary Figure D.2). We confirmed these results by testing multiple crRNAs/dsDNA activators (Supplementary Figure D.3) and Cas12a orthologs, as from the *Francisella novicida* (FnCas12a) and the *Acidaminococcus* (AsCas12a; Supplementary Figure D.4). In addition, the effect of temperature and Mg^{2+} concentration on the LbCas12a-based sensing assay has been investigated because of the synergistic effect on both Cas12a activity and stability of the hairpin DNA structures. The Stem-loop #10T reporter shows again the highest fluorescence increase at the temperatures of interest for diagnostic applications (25 - 45 °C; Figure 5.2c) and optimal $MgCl_2$ concentration is reported to be 10 mM (Figure 5.2d). As expected, at concentration of $MgCl_2$ between 3 and 5 mM, the Stem-loop #10T and #5T reporters generate a lower but similar signal change, probably because the latter structure is more stable and less sensitive to the presence of Mg^{2+} ions.

Of note, at these concentrations of Mg^{2+} the *trans*-activity is still detectable whereas the linear-ssDNA reporter does not provide any significant signal change. To properly compare *trans*-cleavage, we thus selected the experimental condition which is optimal for all the different FRET-based reporters.

Motivated by the enhanced signal transduction obtained using hairpin reporters, we sought to further explore the effect of DNA sequence content on the cleavage activity since previous publications reported a preferential *trans*-cleavage for AT-rich DNA reporters.[61,219] Thus, we tested two variants of the Stem-loop #10T reporter containing different number of thymine in the loop region (Variant #1, loop composed of 12T; Variant #2, loop of 7T+5A). Both the hairpin Variants confirm to better transduce the fluorescence signal compared to the Linear-ssDNA reporter (Supplementary Figure D.5). This suggests that Cas12a *trans*-activity has no or very little preference for primary sequence content of the DNA reporter, and that probably the observed enhancement can be mainly ascribed to the presence of the secondary structure. However, it is very important to distinguish whether improved signal change from background can be ascribed to the real preference of Cas12a for hairpin DNA substrates, or to the intrinsic superior quenching efficiency of the FRET pair in this structure.

5.4.3 Hairpin Reporters Show Enhanced Rate of *Trans*-Cleavage

Motivated by the above considerations, we first carried out time-dependent gel electrophoresis assays of *trans*-activity on the DNA reporters. The *trans*-cleavage activity of the LbCas12a protein is a multi-turnover enzymatic reaction catalysed by the dsDNA target-activated ribonucleoprotein complex that degrades ssDNA probes in a RuvC-

domain-dependent manner. To study the kinetics of LbCas12a *trans*-cleavage activity, we performed direct dsDNA detection without any pre-amplification step by preparing a 10X pre-incubated LbCas12a enzyme/crRNA/ds-DNA target complex (200 nM, 200 nM and 10 nM, respectively) and then diluting the mix 10 times into a solution containing the DNA reporters (100 nM). This is necessary to assume that all target molecules have reacted and that the *trans*-cleavage reaction is the rate-limiting process. The results show faster *trans*-cleavage for hairpin probes compared to the linear one (Figure 5.3a). We quantified the PAGE data by FAM imaging in order to determine the cleavage fraction-time plot (see Table D.1) and then compared it to spectrofluorometric data. Fluorescence data are in agreement with gel assays confirming that all the three hairpin DNA reporters are degraded faster than the Linear-ssDNA reporter, regardless of whether LbCas12a is activated by dsDNA or ssDNA target sequences (Figure 5.3b and Supplementary Figure D.6). We also carried out control experiments to evaluate if a-specific interactions between inactive RNP complex and reporters can affect the cleavage assays. Data obtained at saturating concentration of dsDNA activator (20 nM, equimolar to RNP complex) confirmed that a-specific interactions do not affect the assays (Supplementary Figure D.7).

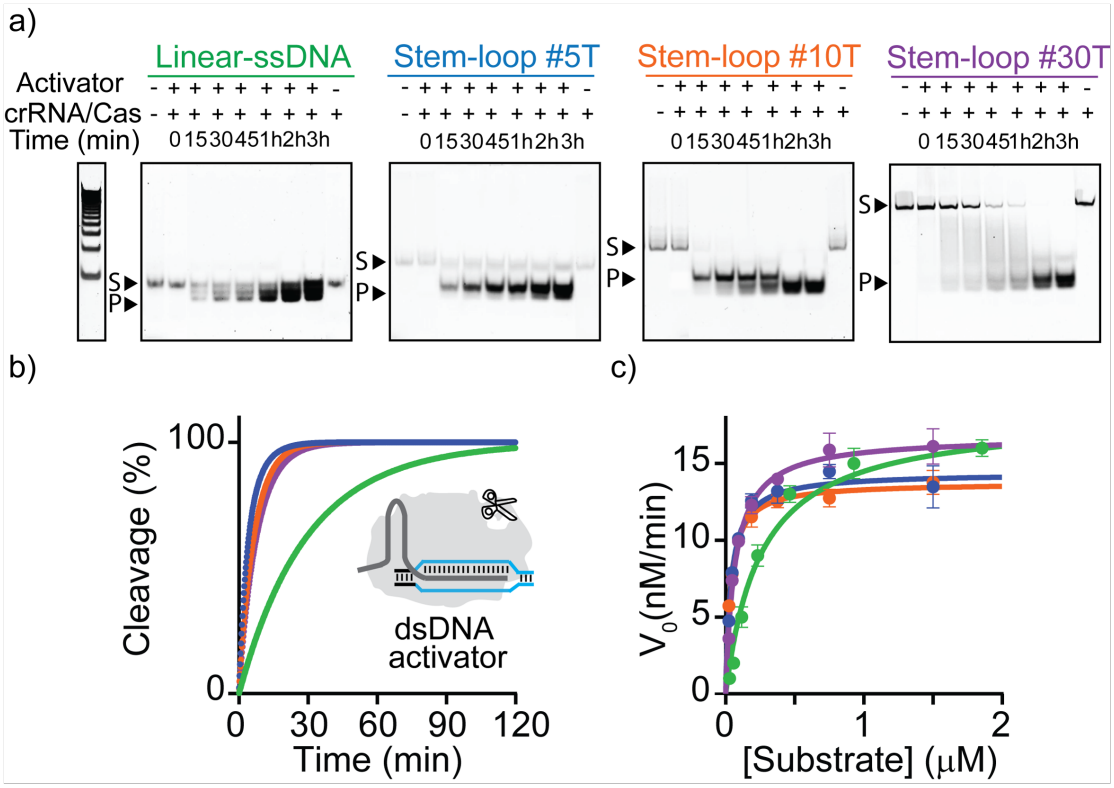


Figure 5.3: **Kinetic analysis of LbCas12a *trans*-cleavage activity.** **(a)** Time-dependent *trans*-cleavage activity of activated LbCas12a on FRET-labelled DNA reporters (250 nM). Substrate (e.g. DNA reporters, S) and degraded products (P) were resolved by polyacrylamide gel electrophoresis (PAGE). The intrinsic fluorescence background of the DNA reporters (lines 1) and signal change over time is consistent with the increased amount of degraded oligonucleotides over time (lines 3 to 8). **(b)** Fraction of DNA reporters cleaved by the active LbCas12a/crRNA/dsDNA complex (RNP complex, final concentration 20 nM/20 nM/1 nM) vs time. Fluorescence assays were performed at 37°C by adding 4.5 μl of a 10x concentrated RNP complex to a solution containing the DNA reporter (final concentration 100 nM). **(c)** Representative Michaelis-Menten plots for LbCas12a *trans*-cleavage activity on the different linear/hairpin DNA reporters using a dsDNA activator. Measured K_m, k_{cat}, and k_{cat}/K_m values reported as mean ± SD, where n = 3.

To study the LbCas12a enhanced activity, we performed a Michaelis-Menten kinetic study of the different DNA reporters. Different concentrations of DNA reporters were employed to determine the initial reaction velocities (V_0, Supplementary Figure D.8). Calibration curves were used to convert V_0 from u.a./min to nM/min (see Supplementary Figure D.8 and Material and Methods Section for the procedure). By

plotting the measured reaction velocities versus DNA reporter (substrate) concentration and fitting the data using Michaelis-Menten equation we obtained k_{cat} and K_m values. The ratio between k_{cat}/K_m is approximately 5-fold higher for the hairpin DNA reporters than for the Linear-ssDNA (Table D.2). This difference can be ascribed to the decrease of K_m values obtained for the three hairpin reporters ($K_{m,\,Linear-ssDNA}$ = 250 ± 50 nM; $K_{m,\,Stem-loop\,\#5T}$ = 42 ± 5 nM; $K_{m,\,Stem-loop\#10T}$ = 34 ± 2 nM; $K_{m,\,Stem-loop\,\#30T}$ = 69 ± 5 nM) whereas k_{cat} values are quite similar ($k_{cat,\,Linear-ssDNA}$ = 0.30 ± 0.02 s^{-1}; $k_{cat,\,Stem-loop\,\#5T}$ = 0.24 ± 0.02 s^{-1}; $k_{cat,\,Stem-loop\,\#10T}$ = 0.23 ± 0.01 s^{-1}; $k_{cat,\,Stem-loop\,\#30T}$ = 0.28 ± 0.02 s^{-1}). These data suggest that the enhanced binding affinity of LbCas12a for hairpin-forming substrates results in enhanced *trans*-cleavage as it is also confirmed through the comparison with previously reported systems (Table D.3). To exclude the possibility that enhanced binding affinity of stem-loop reporters can be due to an a-specific effect generated by the extra number of bases present in the sequence, we performed control assays using a 22 base long DNA reporter (i.e. Variant-ssDNA reporter). It consists of single-stranded DNA terminally labelled with FAM/BHQ1 pair that presents the same sequence of the loop (i.e. 10 T, in the middle of the sequence) of the Stem-loop #10T reporter and additional not self-complementary nucleotides at the 3'- and 5'-end in place of the stem in order to hinder any hairpin-like folding. PAGE analysis of *trans*-cleavage activity over time clearly indicates slower degradation of the Variant-ssDNA reporter and, as expected, lower fluorescence signal transduction (Supplementary Figure D.9) confirming the importance of the hairpin conformation to achieve enhanced *trans*-activity.

5.4.4 Mechanistic Insight into the Hairpin DNA reporter interactions with Cas12a

To elucidate the molecular basis for enhanced *trans*-cleavage, we performed molecular dynamics (MD) simulations and molecular docking studies. Molecular simulations were based on the X-ray structure of Cas12a obtained post non-target strand cleavage (PDB: 5NFV).[2] This structure contains a short dsDNA at the PAM distal site, which should be released from the protein after target strand cleavage to enable the access of ssDNA for *trans*-cleavage.[16] We have removed this portion of dsDNA from the X-ray structure 5FNV, obtaining a state of Cas12a post cis-cleavage and thereby prone for *trans*-cleavage. This system has been an object of MD simulations, collecting 4 μs for two replicates. These simulations revealed a remarkable opening of the system (Figure 5.4a) that lead to the widening of the DNA-binding groove in the RuvC domain. Indeed, the Nuc domain significantly opens, as shown by an increase in the distance between the Rec2 and Nuc domains and an outward bending of Nuc, observed with respect to the RuvC domain (Figure 5.4b). This opens the room for the incoming ssDNA to bind for *trans*-cleavage.[59] Notably, the observed opening of Nuc with respect to REC2 is in line with previous biophysical studies indicating concerted dynamics and high flexibility of the two domains.[1,113] This open Nuc conformation provided us a suitable starting point to perform molecular docking of hairpin and linear DNA structures at the RuvC catalytic site. We selected the sequence of the Stem-loop #5T as our model hairpin structure. First, extensive MD simulations of the Stem-loop #5T and the Linear-ssDNA have been performed (i.e., 200 ns for each system in two replicates) to obtain an ensemble of

conformations for molecular docking. Then, three representative conformations from the top three clusters were selected for molecular docking. Rigid body docking of the DNA models within the RuvC catalytic site was performed using HADDOCK2.4.[230]

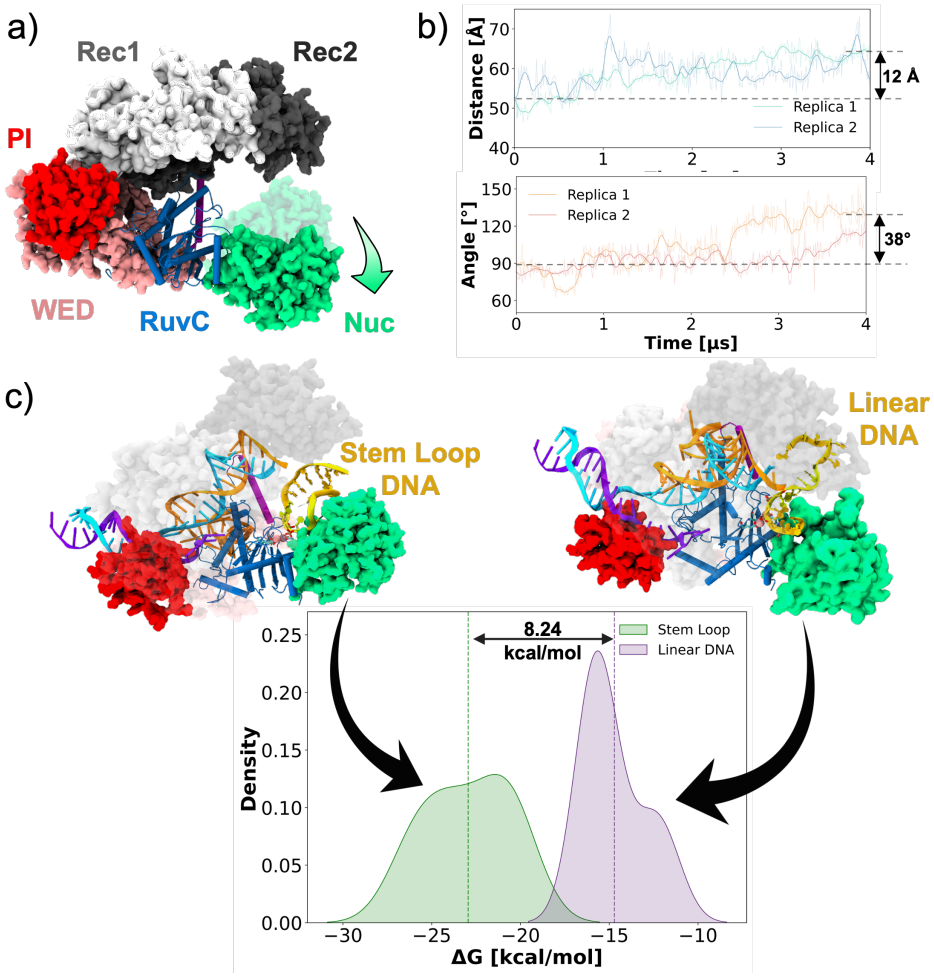


Figure 5.4: Mechanistic insights from molecular simulations. **(a)** Long timescale molecular dynamics (MD) simulations reporting a large-scale opening of the Nuc (indicated by the arrow in green). **(b)** Outward bending of Nuc along ~4 μs MD trajectories (in two replicates), shown through the time evolution of the center of mass distance between the REC2 and Nuc domains (top), and the angle of Nuc with respect to the RuvC domain (bottom). **(c)** Comparison of the binding free energy between the association of the stem loop (green) and linear ssDNA (purple) with Cas12a using the MM-GBSA solvation method.

The most representative docked conformations, based on the electrostatic and Van der Waal energies, and HADDOCK score, were selected for further molecular simulations

and analyses. This approach using MD before and after molecular docking allowed us to reduce the possible biases arising from the docking calculations of single stranded DNA.[243] To compute the binding free energy, 200 ns of unconstrained MD simulations were carried out for both the stem-loop DNA and linear DNA in complex with Cas12a. This was obtained using the Molecular Mechanics with Generalized Born and Surface Area (MM-GBSA) solvation method.[229] This method gave us an appreciable comparison of the binding energies between the two systems. We observed a difference in free energy of 8.24 kcal/mol in favor of the Stem-loop #5T, compared to the Linear-ssDNA (Figure 5.4c and Supplementary Figure D.10). This suggests that the Hairpin DNA model promotes stronger binding with the Cas12a catalytic pocket. Analysis of the interactions further reveals that the stem loop DNA interacts more extensively with the Nuc domain than the linear DNA (Supplementary Figure D.11). Accordingly, the stem loop DNA binds more stably within the Cas12a core, as an effect of increased interactions with Nuc, when compared to the linear DNA. A more stable location of the substrate within the active site can favor the catalysis, ultimately resulting in an increase in the catalytic efficiency. This is in agreement with our experimental observations, reporting a lower K_m and higher cleavage activity for the hairpin DNA reporter when compared with Linear-ssDNA. Moreover, these simulations provide a mechanistic understanding of the binding of hairpin and linear DNA. These findings also take us a step further toward understanding the binding of single stranded DNA (both hairpin and linear) at Cas12a for *trans*-cleavage.

5.4.5 Enhanced Sensitivity and Specificity using Hairpin DNA Reporter

To better characterize the advantages of using a hairpin DNA reporter as fluorescence reporter, we selected the best-performing FRET-based DNA reporter (i.e. Stem-loop #10T) and compared its analytical features with those of the Linear-ssDNA. To do this, we designed an LbCas12a-based detection assays to detect synthetically generated ds-DNA amplicons of the S gene of SARS-CoV-2.[244] Although no clinical samples were tested, we performed quantitative detection of synthetic SARS-CoV-2 dsDNA amplicons (102 bp). The results clearly indicate that the Stem-loop #10T reporter consistently has a higher sensitivity ($slope_{Stem-loop\,\#10T}$ = 137.0 a.u./nM) and lower limit of detection (LOD = 10 pM) for detecting SARS-CoV-2 as compared to the Linear-ssDNA reporter ($slope_{Linear-ssDNA}$ = 61.3 a.u./nM; LOD = 150 pM; Figure 5.5a), and exhibit an average of 2-fold signal improvement between 1 nM and 10 pM of target SARS-CoV-2 dsDNA amplicons. In addition, when samples at known concentration (from 20 aM to 200 pM, see Figure 5.5b) were pre-amplified by PCR (30 cycles, two-pot PCR-CRISPR reaction), the Stem-loop #10T demonstrated higher sensitivity intended as the upstroke of the input/output response in a two-pot reaction assay (i.e. by diluting PCR samples 500 times before the incubation with the LbCas12a/crRNA complex). To demonstrate the generalizability of the results for the detection of bacteria strains, the amplification of nucleotide sequences within the *invA* gene of *Salmonella typhimurium* was evaluated as a means to detect Salmonella serotypes. Salmonella is an important food-borne pathogen responsible for human disease outbreaks worldwide. The reference

method for its detection in food requires up to 5-6 days to produce final results, as it involves pre-enrichment, selective enrichment, isolation on selective agar, biochemical characterisation of presumptive colonies and final serological confirmation.[245] Therefore, there is a need for new methods capable of detecting small amounts of salmonella in food samples in a short time. To this end, several molecular methods for the detection of Salmonella in food have been developed[246,247] and we also report here preliminary results for an alternative and more sensitive detection of this pathogen. The specific PCR product was a 244 bp DNA amplicon which was analysed in CRIPSR-Cas12a assays through serial dilution obtained from the same PCR product. Both the kinetic analysis and the fold change associated to different dilution factors clearly confirmed that enhanced sensitivity and more rapid signal transduction using the stem-loop #10T reporter (Figure 5c and d).

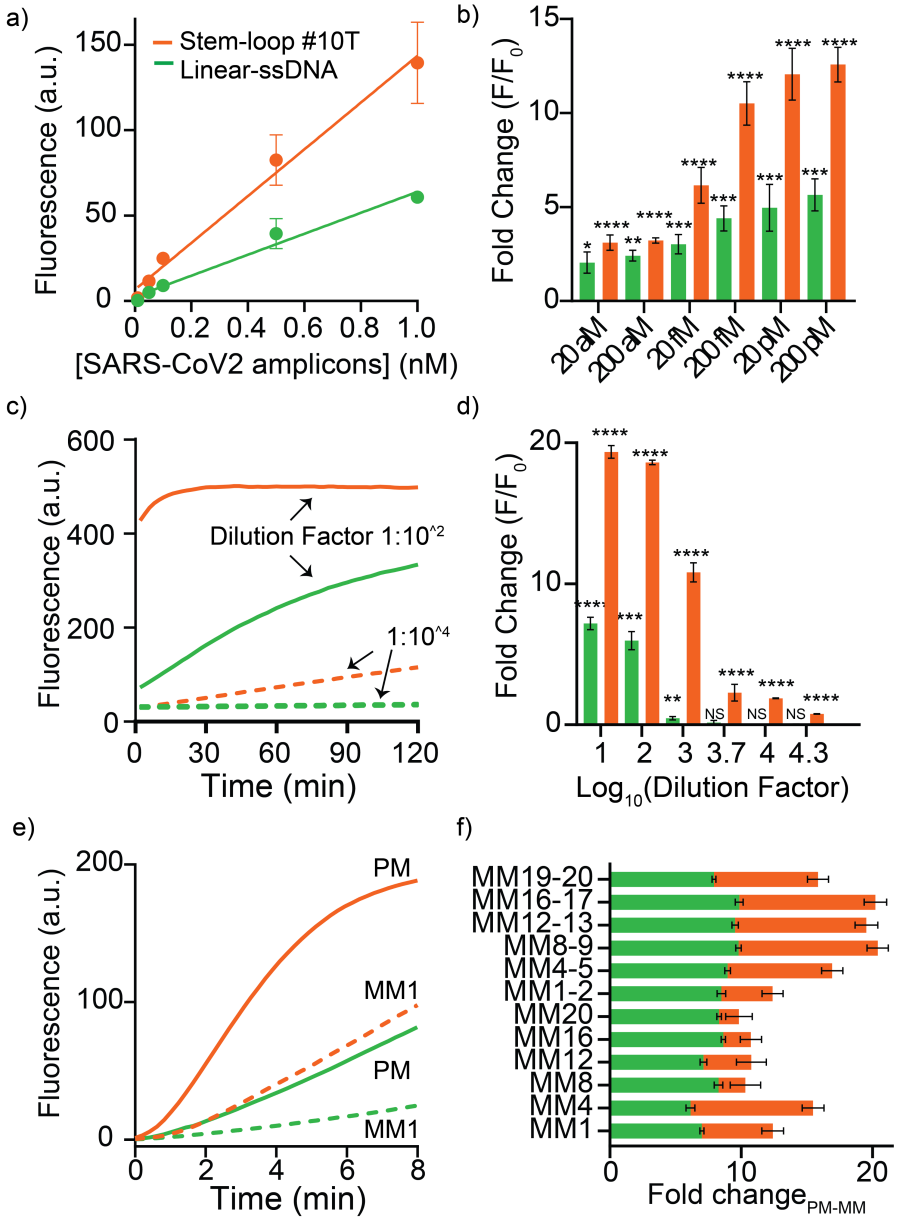

a)
Stem-loop #10T
Linear-ssDNA
Fluorescence (a.u.)
[SARS-CoV2 amplicons] (nM)
b)
Fold Change (F/F0)
20 aM
200 aM
20 fM
200 fM
20 pM
200 pM
c)
Fluorescence (a.u.)
Dilution Factor 1:10^2
1:10^4
Time (min)
d)
Fold Change (F/F0)
NS
Log10(Dilution Factor)
e)
Fluorescence (a.u.)
PM
MM1
Time (min)
f)
MM19-20
MM16-17
MM12-13
MM8-9
MM4-5
MM1-2
MM20
MM16
MM12
MM8
MM4
MM1
Fold changePM-MM

Figure 5.5: **Sensitivity and specificity assays.** **(a)** Amplification free, direct detection of SARS-CoV-2 ds-DNA amplicons. Calibration curves obtained after 1h of incubation of different concentrations of ds-DNA amplicons to a solution containing LbCas12a/crRNA (20 nM, final concentration) and the Linear-ssDNA or the Stem-loop #10T DNA reporter (100 nM). All the values reported show the mean ± SD, where n = 3 replicates. **(b)** Two-pot CRISPR-based detection of PCR amplified SARS-CoV-2 ds-DNA amplicons at known concentrations showing enhanced signal transduction using the Stem-loop #10T reporter. Pairwise comparisons of reactions containing target ds-DNA amplicons with the control (0 copies per μl of ds-DNA amplicon) were done using an unpaired t-test with Welch's correction (n = 5). The asterisks denote values levels of statistical significance: $p < 0.0001$ (****); $p < 0.001$ (***); $p < 0.001$ (**); $p \leq 0.005$ (*). **(c)** Detection of PCR-amplified *Salmonella typhimurium* samples, realized by adding diluted PCR-amplified *Salmonella typhimurium* samples to a solution containing LbCas12a/crRNA (20 nM, final concentration) and the Linear-ssDNA or the Stem-loop #10T DNA reporter (500 nM). Kinetic assays show enhanced signal transduction for stem-loop #30T reporter testing samples at different dilution factors ($1{:}10^2$ and $1{:}10^4$). **(d)** Fold change of two-pot CRISPR-based detection of PCR-amplified *Salmonella typhimurium* at different dilution factors. Pairwise comparisons of reactions containing target PCR-amplified Salmonella amplicons with the control (0 copies per μl of DNA amplicon) were done using an unpaired t-test with Welch's correction (n = 5). The asterisks denote values levels of statistical significance: $p < 0.0001$ (****); $p < 0.001$ (***); $p < 0.001$ (**); $p \leq 0.005$ (*). Comparisons that were not significant (NS) have a P value of > 0.05 or signal that was lower than the control (0 copies per μl of DNA amplicon). **(e)** Specificity test on single-nucleotide mismatch mutations (from MM1 to MM20) and double-nucleotide mismatch mutations (from MM1-2 to MM19-20) on synthetic SARS-CoV-2 dsDNA activator. Kinetic assays report the fluorescence difference between perfect match (PM) and mismatch 1 (MM1) target (right, t = 4 min). **(f)** Superimposed bar graphs showing the difference between the fold change related to the perfect match (PM) and to the mismatch (MM) target (right, 10 nM at t = 4 min). All the values reported show the mean ± SD, where n = 3 replicates.

We evaluated the specificity of LbCas12a-based detection in discriminating point mutations across dsDNA amplicons using both the Stem-loop #10T and the Linear-ssDNA reporters. We mutated either a single or two consecutive nucleotides at different points of the dsDNA amplicons of SARS-CoV-2. Since Stem-loop #10T reporter show enhanced and faster cleavage rate compared to Linear-ssDNA, we hypothesized that by performing short-time fluorescence analysis of *trans*-cleavage activity (t = 4 minutes), discrimination between perfect match (PM) and mismatch (MM) targets could be improved (Figure 5.5e and Supplementary Figure D.12). As expected, the single-point mutations were more easily tolerated than the double-point mutations by LbCas12a. Nevertheless, it is noteworthy that the difference of fold signal change between PM and MM is significantly higher for the Stem-loop #10T compared to Linear-ssDNA reporter at different ds-DNA target concentration (Figure 5.5f and Supplementary Figure D.13). Cleavage PAGE assays over time using mismatches confirmed that the significant difference of *trans*-cleavage rate between the two reporter systems allows to better discriminate the presence of MMs using Stem-loop #10T (Supplementary Figure D.14).

In this work we have studied a set of hairpin DNA reporters with loop domains of different length as FRET-based reporters of the *trans*-activity of Cas12a enzymes. We have first confirmed that the loop length can be used to finely regulate the switching mechanism between the folded (ON)/unfolded (OFF) state of the reporters, and this also provides a means to rationally control the overall fluorescence signal change produced by the *trans*-activity of Cas12a. Then, kinetic analysis of the *trans*-activity showed that all the hairpin DNA reporters have faster Cas12a-based cleavage rate compared to that observed using short linear ssDNA probes. In addition, the Michaelis-Menten

analysis performed on the different DNA reporters indicated an enhanced affinity (lower K_m values) of Cas12a for the hairpin DNA reporters. Molecular dynamics simulations of Cas12a showed a remarkable opening of the system that lead to the widening of the DNA-binding groove in the RuvC domain, and docking studies confirmed the stronger binding of stem-loop DNA to the Cas12a catalytic pocket. Finally, we tested the best performing hairpin DNA reporter (Stem-loop #10T) for the detection of synthetic ds-DNA amplicons related to S gene of SARS-CoV-2 and Salmonella strains. The use of the Stem-loop #10T resulted in a sensing platform with improved sensitivity and enhanced specificity in discriminating single and double point mutations across dsDNA amplicons. These data suggest that the enhanced affinity of Cas12a for hairpin probes, combined to the well-established capacity of FRET-based stem-loop DNA to improve fluorescence signal transduction, can significantly enhance the sensitivity and the specificity of CRISPR-Cas12-based sensing platforms. The faster signal transduction and the ability to discriminate among different DNA sequences offers the opportunity to use hairpin-based transducer of *trans*-activity in a variety of high-throughput applications, from metagenomics to biomedicine.

Our findings represent a step toward a better understanding of Cas12a nuclease activity that can potentially assist in the development of improved diagnostic platforms for the diagnosis of various diseases, including the COVID-19. Beyond the diagnostic applications, this work paves the way for exploring evolutionary and mechanistic aspects of CRISPR type V enzyme that govern the affinity of CRISPR Type V systems for hairpin-like structures. Further studies are currently in progress to understand whether other CRISPR Type systems (Cas13) showing *trans*-activity show a similar behaviour.

5.5 Acknowledgements

This work was financially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sk lodowska-Curie grant agreement "Nano-Oligo Med" (No 778133), and Italian Ministry of University and Research (Project of National Interest, PRIN, 2017Y2PAB8_004 through the project "Cutting Edge Analytical Chemistry Methodologies and Bio-Tools to Boost Precision Medicine in Hormone-Related Diseases". M. Rossetti was supported from a Fondazione Umberto Veronesi postdoctoral fellowship. Computational studies are based upon work supported by the National Institutes of Health (Grant No. R01GM141329, to GP) and the National Science Foundation (Grant No. CHE-1905374, to G.P). This work used resources, services, and support provided via the COVID-19 HPC Consortium (https://covid19-hpc-consortium.org/), which is a unique private-public effort to bring together government, industry, and academic leaders who are volunteering free compute time and resources in support of COVID-19 research (Grant No. MCB200150, to G.P.). Computer time has also been awarded by XSEDE under Grant No TG-MCB160059 and by NERSC under Grant No M3807 (to G.P.).

Chapter VI

Dimerization of the Deaminase Domain and its Interactions with Cas9 Boost Base Editing Efficiency in ABE8e

Reprinted with permission from Arantes P. R. *et al.*

6.1 Abstract

CRISPR-based DNA Adenine Base Editors (ABEs) hold remarkable promises to address human genetic diseases caused by point mutations. Here, extensive molecular simulations and experimental approaches reveal the biophysical basis behind the astonishing base editing efficiency of ABE8e, the most efficient ABE to date. We show that the dimerization of the TadA8e (the deaminase domain) and its unique juxtaposition to Cas9 are pivotal for the efficient DNA deamination by ABE8e. The TadA8e dimer

forms robust interactions involving its R98 and R129 residues, the RuvC domain of Cas9 and the DNA. These interactions are exclusive to ABE8e, distinguishing it from its predecessor, ABE7.10, and are indispensable for efficient DNA deamination. Three critical substitutions, T111R, D119N and N122H, strikingly balance enhanced activity with stability, reinforcing the TadA8e dimer and improving the ABE complex's functionality. These insights into the ABE8e's molecular mechanisms offer new directions to advance the development of safer precision genome editing tools.

6.2 Introduction

The bacterial immune systems, CRISPR-Cas, have ushered in a new era of genome editing, offering powerful and easily customizable tools for precise genetic manipulation.[248] These tools rely on the special feature of noncoding RNA-guided CRISPR-Cas effector complexes: their ability to introduce double-stranded DNA breaks (DSBs) at specific genomic DNA loci in a programmable manner.[8,10,249] The outcome of the CRISPR-Cas-mediated genome editing depends on which DNA repair pathway addresses the introduced DSB. Homology-directed repair (HDR) produces accurate DNA modifications, but it is inefficient since it is limited to actively dividing cells and requires a donor DNA template.[250]

DNA base editing is one of the newest CRISPR-Cas technologies that allows direct and irreversible conversion of one base pair to another at a specific genomic site without relying on DSBs or HDR.[251] This technology holds tremendous promise in treating a wide range of genetic diseases associated with single nucleotide polymorphisms

(SNPs),[252] due to its ability to precisely convert A•T base pairs to G•C (by adenine base editors, ABEs) and C•G to T•A (by cytosine base editors, CBEs).[253–258] ABEs are chimeric enzymes composed of a catalytically impaired Cas9 [Cas9 nickase (nCas9)] from *S. pyogenes* and a deoxyadenosine deaminase (Figure 6.1).[259] These machineries use the nCas9 – guide RNA complex to locate DNA sequences of interest within the genome, and a deaminase enzyme to perform A-to-G modifications via deamination of deoxyadenosine to deoxyinosine at specific single-stranded DNA (ssDNA) sites.

The deaminase domain of ABEs' was evolved from an *Escherichia coli* transfer RNA (tRNA) adenosine deaminase (wt-TadA), which naturally deaminates adenosine in the anticodon loop of tRNA.[254,260,261] During multiple rounds of directed evolution, mutations were introduced to the wt-TadA, leading to evolved TadA enzymes (TadA*). When conjugated with nCas9, these enzymes demonstrate increasing efficiency in deaminating deoxyadenosine in ssDNA across generations (measured as the number of A•T converted to G•C at different genomic sites).[254,262] Among the ABEs used for precision genome editing in various cells and organisms, ABE8e is one of the most efficient and widely used (with DNA base editing efficiency of 80-90 %).[262,263] In ABE8e, the TadA8e deaminase domain contains a total of 20 mutations with respect to the wt-TadA and is fused to nCas9 likely forming homodimers in *trans* (Figure 6.1).[259,262] Compared to its closest predecessor, ABE7.10, ABE8e exhibits significant advancements. The additional eight mutations that were introduced to ABE7.10 to yield ABE8e not only enhance the efficiency of DNA editing in cells, soaring from approximately 58 % to ∼80-90 %, but also drastically improve the deamination rates (580-fold or more) measured in vitro under single turnover conditions.[259,262]

The cryo-EM structure of ABE8e in a substrate-bound state captured dCas9 (catalytically dead Cas9 with D10A and H840A substitutions) fused with the TadA8e deaminase (Figure 6.1, Supplementary Figure E.2).[259] dCas9-guide RNA complex binds the DNA by matching one strand with the guide RNA (the target strand, TS) while the other non-target strand (NTS) is in a single-stranded state. TadA8e is captured as a dimer with one domain deaminating deoxyadenosine within the DNA NTS ($TadA^{cat}$, red) and another domain docking to the dCas9 protein ($TadA^{dock}$, pink). While this structure provides critical insights into the mechanism of DNA deamination, the biophysical rationale behind the astonishing ABE8e base editing efficiency remains unresolved. Additionally, it is perplexing how the eight amino acid substitutions lead to a remarkably sharp increase in the deamination efficiency of ABE8e compared to its predecessor ABE7.10. Nevertheless, this efficiency comes with elevated off-target editing rates.[262] Thus, understanding the molecular basis of ABE8e's enhanced DNA base deamination is crucial for designing more precise genome editors and accelerating their translation for biomedical applications.

Here, we establish the biophysical basis behind the improved DNA base editing efficiency of the ABE8e enzyme through an integrative approach that combines extensive molecular dynamics (MD) simulations and enhanced sampling approaches with biochemical and biophysical experiments. An overall ensemble comprising multi-microsecond (μs) MD simulations was combined with ensemble Förster Resonance Energy Transfer (FRET) experiments, thermal stability assays, and single turnover kinetics.

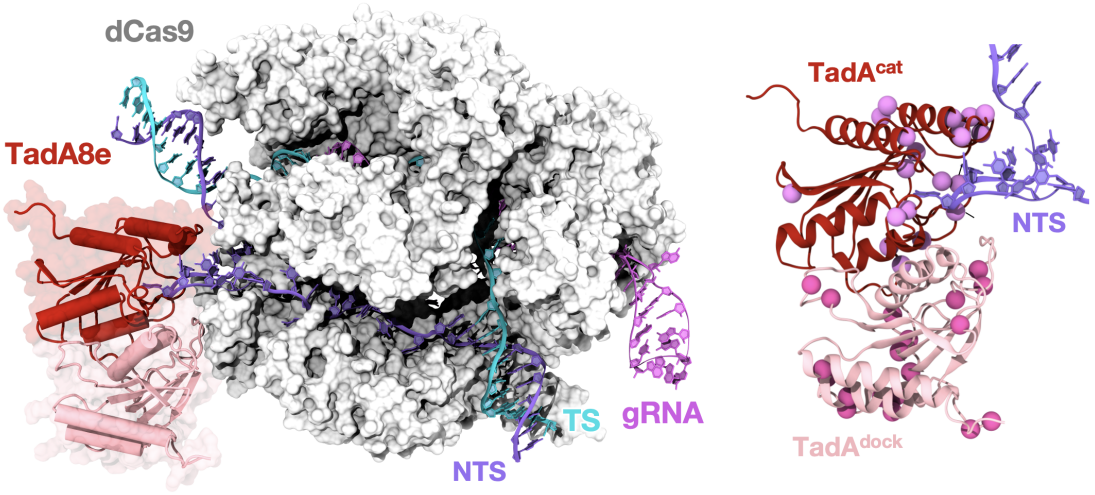


Figure 6.1: **Overview of Adenine Base Editor's (ABE) structure.** Cryo-EM structure of the catalytically dead Cas9 (dCas9) from *S. pyogenes* fused to the adenosine deaminase TadA8e (PDB 6VPC).[259] The fusion of dCas9 and TadA8e (i.e., ABE8e) is shown as a molecular surface. The dCas9 protein binds a guide RNA (gRNA, magenta), unwinds double-stranded DNA, base-pairs the guide RNA (gRNA) with the DNA target strand (TS, cyan) and displaces the other non-target strand (NTS, violet). TadA8e forms a dimer with one domain deaminating deoxyadenosine within the DNA NTS [a catalytic domain, TadA^{cat} (red)] and another domain docking between the TadA8e^{cat} and dCas9 [a docking domain, TadA^{dock} (pink)]. On the left, a close-up view of TadA8e shows an array of amino acid substitutions (spheres, magenta) introduced to the wt-TadA through directed evolution.

We reveal that TadA8e dimerization and its unique juxtaposition to dCas9 are pivotal for the efficient DNA deamination in ABE8e. Indeed, the TadA8e dimer is more stable than that formed by its wt-TadA progenitor, which allows TadA8e to establish crucial interactions with DNA and dCas9, contributing to improved DNA deamination efficiency. Two residues within the TadA8e docking domain, R98 and R129, significantly increase their interactions with the dCas9-DNA complex in comparison to ABE7.10 and are required for efficient DNA deamination. Moving from ABE7.10 to ABE8e, the interplay between two critical substitutions, T111R and D119N, is shown to increase DNA deamination rates and balance the stability of the complex, while H122N contributes to

stabilizing the TadA dimer, ultimately evolving the complex toward improved functionality. These findings hold significant implications for the design of ABE enzymes and open new directions for developing improved base editors.

6.3 Results

6.3.1 Multiple rounds of directed evolution have stabilized the dimeric state of TadA

ABEs have been developed to deaminate DNA by mutating the wt-TadA, a homo-dimeric enzyme with each monomer able to catalyse adenosine deamination in tRNA. The tRNA-binding pocket is formed by both monomers[261,264] suggesting that dimerization might be important for efficient tRNA deamination. To establish whether multiple rounds of directed evolution have impacted the dimerization of TadA, and to characterize its role in DNA adenine base editing, we performed free energy simulations of TadA8e and compared it to wt-TadA. We used a two-dimensional Umbrella Sampling (2D US)[117] method to describe the dimerization process of the TadA enzymes along two reaction coordinates: (*i*) the inter-monomeric distance considering the center of mass of each TadA unit, and (*ii*) the difference in root-mean-square deviation (RMSD) of the Cα atoms with respect to the experimental structures (i.e., PDB 1Z3A[261] for wt-TadA and PDB 6VPC[259] for TadA8e). Combining these reaction coordinates enabled us to describe the dimerization process while keeping the experimental structures as a reference (details in the Methods section). These runs were performed on the TadA dimers in solution, without the dCas9, describing the potential of TadA to dimerize *per se*. For both TadA

variants, we rigorously collected the population states from dimer to monomer (and vice versa) through $\sim$25 μs of converged sampling (Supplementary Figure E.3). The 2D free energy surfaces reveal that both wt-TadA and TadA8e reach a stable minimum that corresponds to the experimental dimeric state (Figure 6.2a), with TadA8e exhibiting a deeper minimum than the wt-TadA. Analysis of the minimum free energy paths (details in the Methods section) also reveals that to escape the minimum of the dimeric state toward the monomer, the wt-TadA is favoured by ¿10 kcal/mol, compared to TadA8e (Figure 6.2b). This indicates that TadA8e is more stable in the dimeric form than its wt-TadA predecessor.

To experimentally assess the dimerization of TadA in the presence of dCas9, we performed FRET experiments. We generated FRET constructs of the ABE0.1 (containing a single wt-TadA domain) and ABE8e (containing a single TadA8e domain), where all cysteines (except C87 and C90 which are present in the active site of TadA) were substituted to serine and T133 of TadA* was mutated to cysteine for site-specific labelling with maleimide derivatives of Cy3 and Cy5 fluorophores (Supplementary Figure E.4a, Supplementary Table E.1). The control labelling reaction of cysteine-free ABE8e confirmed site-specific labelling of the ABE8e FRET construct (Supplementary Figure E.4b). The Cy3/Cy5-labelled ABE0.1 and ABE8e are properly folded (Supplementary Figure E.4c), and the dye-labelling does not affect the deamination rates of ABE0.1 and ABE8e (Supplementary Figure E.4d-g). Our ensemble FRET experiment shows that both wt-TadA and TadA8e enzymes exist in dimeric form when fused to dCas9 (similar to what was observed for the enzymes in solution through free energy simulations), as evidenced by the presence of a peak at $\sim$675 nm in the fluorescence emission spectra

of Cy3/Cy5-labelled proteins when excited at 530 nm. To compare the dimerization affinity of wt-TadA and TadA8e, we conducted FRET measurements on dye-labelled samples at varying protein concentrations (Figure 6.2c). These measurements show that TadA8e forms highly stable dimers, with K_d values below 4 nM (likely in the picomolar range). In contrast, the wt-TadA has a propensity to dissociate, suggesting a higher K_d compared to TadA8e (likely in the nanomolar range), as evidenced by the decrease of Cy5 fluorescence emission intensity as the enzyme concentration decreases. Taken together, these findings suggest that multiple rounds of directed evolution have stabilized the dimeric state of TadA8e, which could critically contribute to their increasing base editing efficiency.

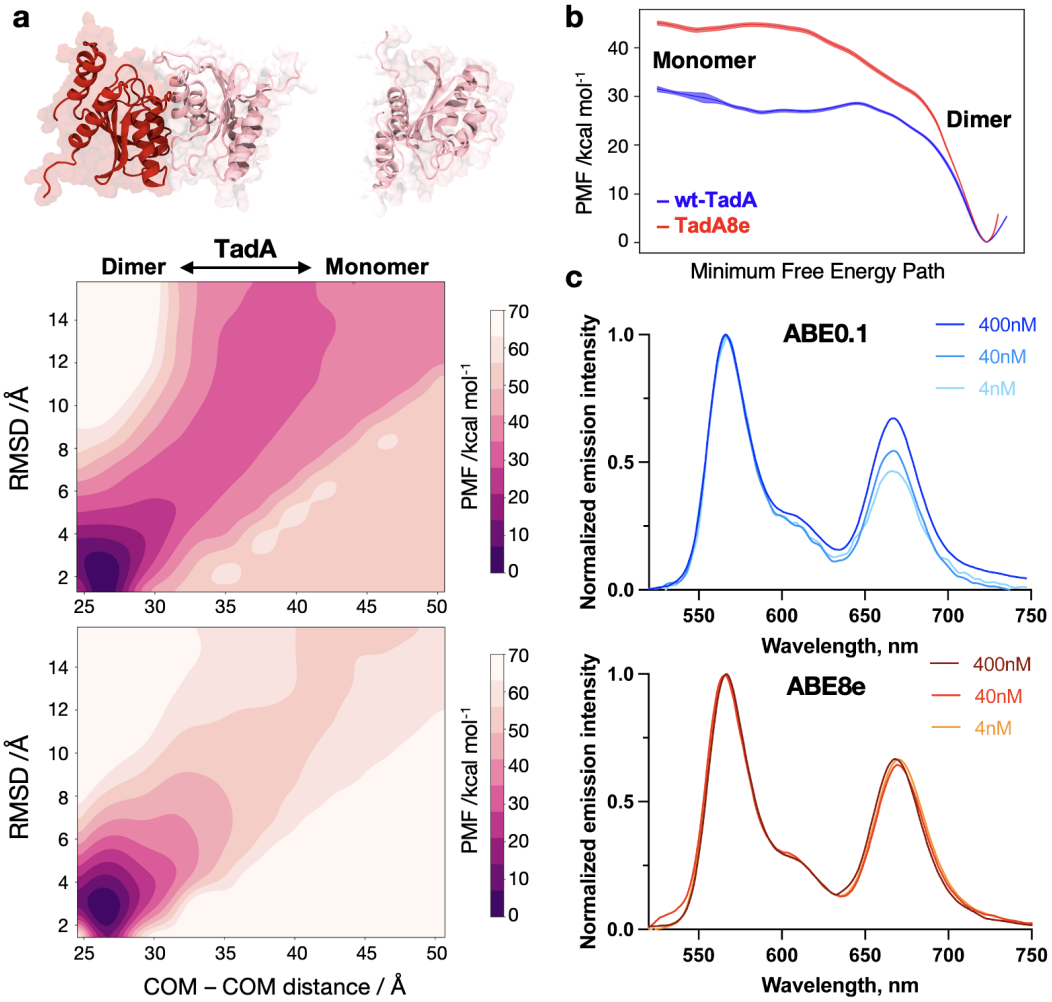


Figure 6.2: **Dimerization of TadA enzymes.** **(a)** Potential of TadA enzymes to dimerize in solution. The potential of mean force (PMF, in kcal mol-1) for the dimerization of the wt-TadA (top) and TadA8e (bottom) enzymes is computed through 2-D Umbrella Sampling simulations along (*i*) the inter-monomeric distance considering the center of mass of each TadA unit, and (*ii*) the root-mean-square deviation of the Cα atoms between the monomeric and dimeric states. The dimer and monomer states of TadA enzymes are indicated and shown on top of the plots. **(b)** Minimum free energy paths for the transition of the wt-TadA and TadA8e enzymes from monomer to dimer, computed from 2-D Umbrella Sampling simulations (see the Methods section). **(c)** Fluorescence emission spectra of Cy3/Cy5-labeled ABE0.1 (with single wt-TadA, top) and ABE8e (with single TadA8e, bottom) across varying concentrations, normalized to the maximum Cy3 donor peak at 570 nm to eliminate the effect of sample dilution. The distinct reduction of FRET signal observed exclusively in the spectra of ABE0.1, as protein concentration decreases, provides a clear indication that TadA8e has a higher affinity for dimerization compared to wt-TadA.

6.3.2 DNA–TadA–Cas9 interactions boost DNA deoxyadenosine deamination by ABE8e

To better understand the importance of TadA dimerization when fused to dCas9, and to characterize the interactions that the TadA enzymes establish with Cas9 and DNA, we performed extensive multi-μs MD simulations of the ABE complexes. We considered ABE0.1 (where TadA is WT), ABE8e, and ABE7.10, the closest predecessor of ABE8e with $\sim$580-fold lower DNA deamination rate.[259,262] The cryo-EM structure of the ABE8e complex (PDB 6VPC),[259] containing dCas9 and the TadA8e homodimer, was used as a template, as it preserves the *bona fide* configuration of the TadA enzymes fused to dCas9. The ABE0.1, ABE8e, and ABE7.10 model systems, comprising $\sim$500k atoms, were simulated for $\sim$9 μs each (in three replicates of 3 μs). An in-depth analysis of our all-atom MD simulations was carried out to describe the interactions established between the TadA domains themselves, as well as those that the TadA enzymes form with dCas9 and the DNA. We computed the pairwise interactions among residues in the ABE complexes using the MM/GBSA method,[229,265,266] assessing the relative interaction energies between different systems, and comparing the gain/loss in interaction strength (details in the Supplementary Methods).

We first compared the ABE0.1 and ABE8e systems, using Sankey plots to report the difference in pairwise interaction energy (ΔE) between them, computed at the triad of interfaces between the TadA units, dCas9, and the DNA NTS (Figure 6.3a). In these plots, residues are connected through edges, whose width is proportional to ΔE. Interactions that gain strength in ABE8e ($\Delta E < 0$, blue) are compared to those that

gain strength in ABE0.1 ($\Delta E > 0$, red). This analysis reveals an increased number of interactions that gain strength at the TadA dimeric interface in TadA8e, compared to the wt-TadA (Figure 6.3a, Supplementary Figure E.5). This is in line with our ensemble FRET measurements indicating that TadA8e forms more stable dimers than wt-TadA (Figure 6.2c). Interestingly, ABE8e also shows a remarkable increase in interactions that gain strength at the interface of the TadA8e docking domain and the DNA NTS, as well as the TadA8e docking domain and the RuvC domain of Cas9, with respect to ABE0.1. This underscores that ABE8e could strengthen the interactions at these interfaces as a possible strategy for improved catalytic efficiency.

We then compared ABE8e with ABE7.10, the closest predecessor of ABE8e with $\sim$580-fold lower DNA deamination rate[259,262] (Figure 6.3b, Supplementary Figure E.6). We observed that, while the number of interactions that gain strength at the TadA dimeric interface is comparable, ABE8e remarkably increases the number of interactions that gain strength between the TadA docking domain, the DNA NTS and the RuvC domain of dCas9, compared to ABE7.10. In Figure 6.3c, the relative pairwise ΔE between ABE8e and ABE7.10 are plotted on the 3D structure at the TadA–dCas9–DNA triad of interfaces showing a noteworthy gain in interaction strength between the TadA docking domain, the two antiparallel β-sheets of the nCas9 RuvC domain (residues 1045 to 1062), and the DNA NTS. This may not only be attributed to the additional eight mutations in TadA8e but also to the fact that TadA8e forms a homodimer, while TadA7.10 is a heterodimer of an evolved TadA7.10 and the wt-TadA. We thereby conducted a comparison of the interaction gain/loss between the homodimeric and heterodimeric TadA8e bound to CRISPR-Cas9 (i.e., ABE8e vs. ABE8e hetero) based on additional

$\sim$9 μs of MD simulations of ABE8e including the TadA8e heterodimer. We observed that the homodimeric TadA8e maintains an elevated number of interactions with dCas9 and the DNA that gain strength, compared with its heterodimeric form (Supplementary Figure E.7). This indicates that the homodimeric TadA8e establishes a strong network of interactions between its docking domain, the two antiparallel β-sheets of RuvC, and the DNA NTS that solidify the juxtaposition of TadA8e with the Cas9-DNA complex as a unique feature of ABE8e.

The Sankey diagrams comparing ABE8e and ABE7.10 (Figure 6.3b) pinpoint that R129 remarkably strengthens the interactions with several residues of the RuvC domain in ABE8e compared to ABE7.10. Concomitantly, R98 binds multiple bases of the NTS while also interacting with dCas9 E1053.

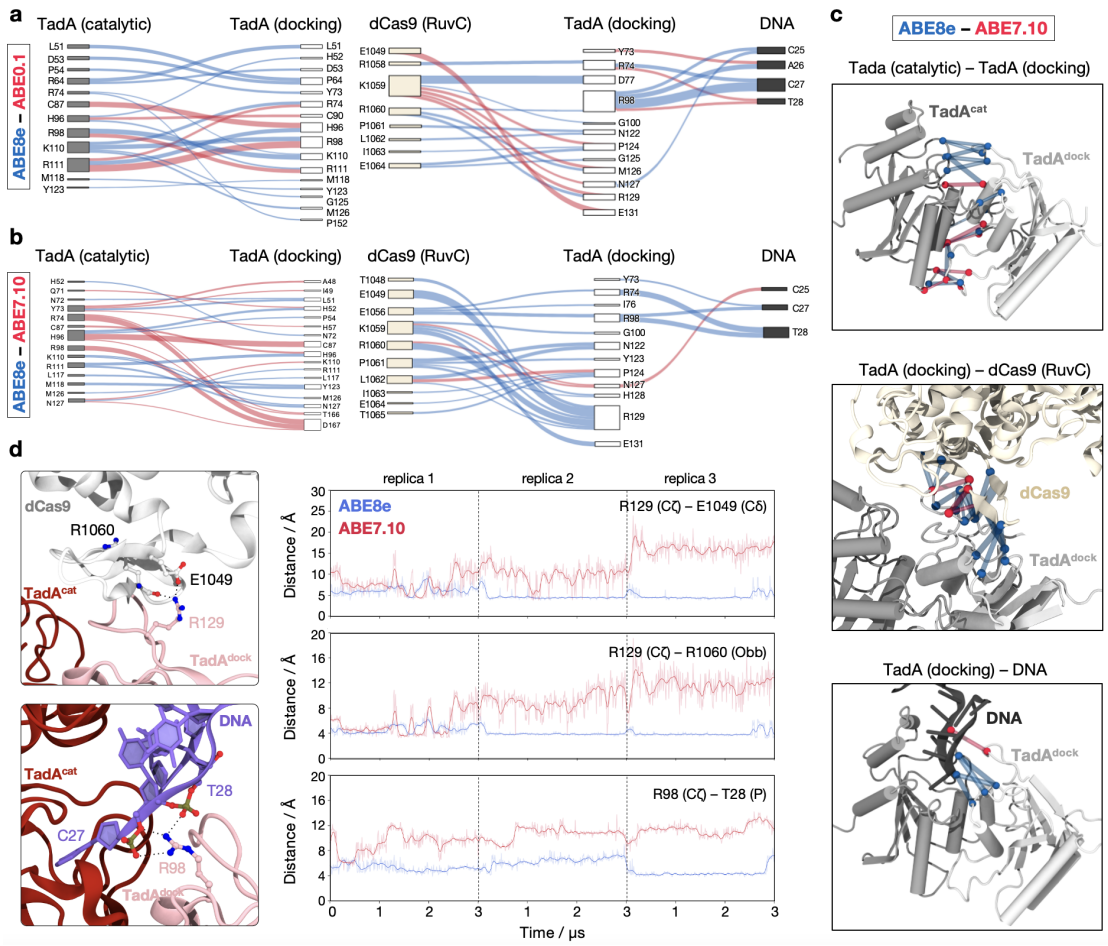


Figure 6.3: **Gain/loss of interaction strength in ABE8e compared with ABE0.1 and ABE7.10. (a-b)** Sankey diagrams of the relative change in pairwise interaction energy (ΔE) between ABE8e and ABE0.1 (a), and between ABE8e and ABE7.10 (b), computed at the interfaces between the TadA catalytic and docking domains (left), the TadA docking domain, the dCas9 RuvC domain, and the DNA non-target strand (right). Sankey diagrams connect residues through lines the thickness of which is proportional to ΔE. Interactions that gain strength in ABE8e (ΔE¡0, blue) are compared to those that gain strength in ABE0.1 and ABE7.10 (ΔE¿0, red). A cut-off of 0.7 kcal/mol was used to filter interactions with substantial perturbation. Details are reported in the Supplementary Methods. **(c)** The relative pairwise ΔE between ABE8e and ABE7.10 are plotted on the three-dimensional structure of the complex and shown at the TadA–Cas9–DNA interface. Interactions that gain strength in ABE8e (blue) are compared to those that gain strength in ABE7.10 (red). Bonds are proportional to ΔE. **(d)** Time-evolution of the interactions established by R129 and R98 with the RuvC domain of dCas9 and the DNA non-target strand along MD simulations of the ABE8e (blue) and ABE7.10 (red) complexes. Data are reported for three simulation replicates of $\sim$3 μs each.

A time-evolution analysis of the interactions shows that in ABE8e, R129 unequivocally interacts with the E1049 side chain and the R1060 backbone carbonyl of dCas9, while these interactions are comparatively less stable in ABE7.10 (Figure 6.3d). R98 binds the NTS phosphate backbone by persistently interacting with thymine 8 (T28) in ABE8e, while this interaction is mostly absent in ABE7.10. Interestingly, R98 loosely binds E1056 of dCas9 in both ABE8e and ABE7.10 (Supplementary Figure E.8), suggesting that R98 mainly binds to DNA in ABE8e. Taken together, the TadA8e docking domain notably strengthens its interactions with both Cas9 and the DNA NTS, suggesting that it could critically contribute to the 580-fold increased DNA deamination efficiency in ABE8e compared to ABE7.10.

We thereby performed a mutational analysis and evaluated the significance of the TadA8e docking domain (R98 and R129) interactions with the DNA and Cas9 for *in vitro* DNA deoxyadenosine deamination. This is accomplished by first designing a construct that contains two TadA8e units (the unit adjacent to the Cas9 construct is regarded as the C-terminal domain of TadA8e, and the other unit constitutes the N-terminal domain) fused to dCas9. Each TadA8e is encoded by different DNA sequences while retaining the same amino acid sequences, which allows us to introduce mutations specifically to only one TadA8e unit at a time. To distinguish which TadA8e acts as a catalytic unit, we inactivated each TadA8e one at a time by substituting the catalytic E59 with A (Supplementary Figure E.9a). DNA deoxyadenosine deamination kinetics of these ABE8e variants in single-turnover conditions yielded identical observed deamination rates (k_{obs}), which are half of the rate observed for ABE8e* when both TadA8e units

were active (Supplementary Figure E.9b, c). This indicates that either of the TadA8e domains can deaminate deoxyadenosine within the NTS.

Following the original assumption in the ABE design that the C-terminal TadA domain is catalytic8, and to match the constructs commonly used in the genome engineering and our computational model (Figure 6.1),[259] we opted for the ABE8e*A construct, where the N-terminal domain is the docking unit and the C-terminal is the catalytic unit (Figure 6.4a, Supplementary Figure E.9a). We then generated two mutants with the R98A substitution in either the TadA8e catalytic domain [ABE8e*A (R98A^{cat})] or the docking domain [ABE8e*A (R98A^{dock})] and measured their DNA deoxyadenosine deamination kinetics under single-turnover conditions (Figure 6.4b, Supplementary Figure E.10a). In agreement with our computational analysis, R98A substitution in the TadA8e docking domain significantly reduced the observed DNA deamination rate, whereas the same substitution in the catalytic unit did not affect the DNA deamination rate (Figure 6.4b). The CD spectrum of ABE8e*A (R98A^{dock}) aligns well with that of ABE8e*A (Supplementary Figure E.10b), indicating that the decrease in DNA deamination activity is attributable to the disruption of R98 interactions with the DNA NTS and/or the RuvC domain of dCas9, rather than protein misfolding. This supports our hypothesis that the docking unit of TadA8e contributes to the DNA deamination rate of the catalytic TadA8e. To assess whether R98 in the docking TadA8e interacts with the DNA or Cas9, we created two versions of ABE8e*A, introducing the R98E mutation in the docking TadA8e [ABE8e*A (R98E^{dock})] (to disrupt TadA8e-Cas9 RuvC interaction) and the same substitution in the docking TadA8e with the complementary

E1056R mutation in dCas9 [ABE8e*A (R98E^{dock}/E1056R^{Cas9})] (to reinstitute TadA8e-Cas9 RuvC interaction). Single-turnover kinetics showed that the DNA deamination rates and efficiencies were similar for both mutants (Supplementary Figure E.10b-d), supporting our computational observation that R98 in the docking TadA8e primarily binds to the NTS of DNA instead of forming strong interactions with Cas9 (Figure 6.3, Supplementary Figure E.8c).

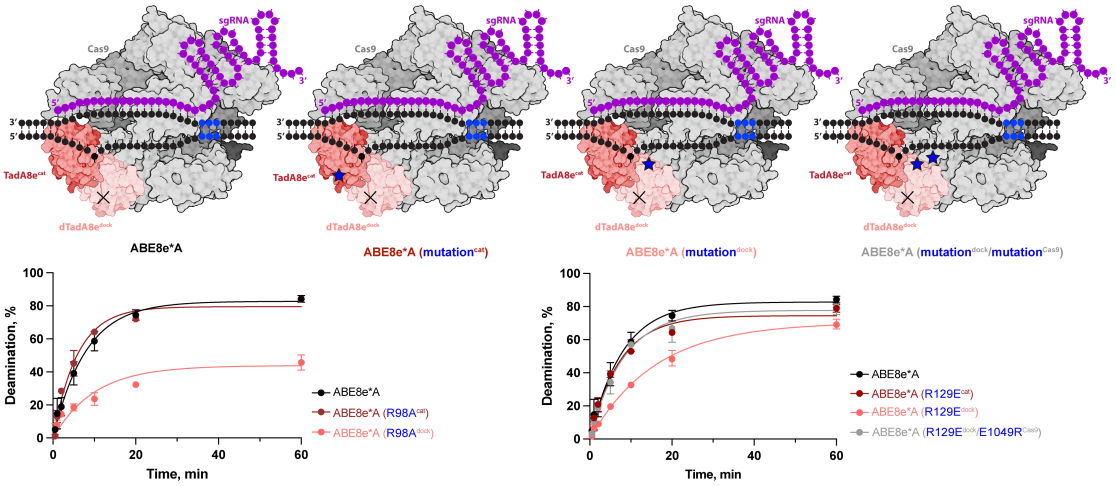

sgRNA
Cas9
TadA8e^cat
dTadA8e^dock
ABE8e*A
ABE8e*A (mutation^cat)
ABE8e*A (mutation^dock)
ABE8e*A (mutation^dock/mutation^Cas9)
Deamination, %
Time, min
ABE8e*A
ABE8e*A (R98A^cat)
ABE8e*A (R98A^dock)
ABE8e*A (R129E^cat)
ABE8e*A (R129E^dock)
ABE8e*A (R129E^dock/E1049R^Cas9)

Figure 6.4: **Key interactions between the docking domain of TadA8e with the DNA and Cas9 for efficient DNA deoxyadenosine deamination by ABE8e. (a)** Model depicting the chosen construct, ABE8e*A, alongside the three different types of mutants designed to evaluate the role of the docking unit of TadA8e in efficient DNA deamination. The catalytic domain of TadA8e (TadA8e^{cat}) is coloured in dark red; the docking domain (TadA8e^{dock}) is pink. A blue star indicates the position of amino acid substitutions introduced to either the catalytic or docking TadA8e domains and the Cas9 RuvC domain. **(b)** Single-turnover kinetics of ABE8e*A and its variants with the R98A substitution introduced to either the catalytic (red) or the docking (pink) domains measured using fluorescently labelled dsDNA containing a single adenine within the editing window. The DNA deamination efficiency and rate of the ABE8e*A (R98A^{dock}) ($k_{obs} = 0.03 \pm 0.01$ min^{-1}) are significantly lower than those of the ABE8e*A ($k_{obs} = 0.12 \pm 0.01$ min^{-1}) and the ABE8e*A (R98A^{cat}) ($k_{obs} = 0.17 \pm 0.02$ min^{-1}) (data analysed with one-tailed t-tests, and p = 0.0001). **(c)** Single-turnover kinetics of ABE8e*A and its variants with R129E substitution introduced to either the catalytic (red) or the docking (pink) domains measured using fluorescently labelled dsDNA containing a single adenine within the editing window. The observed rates of DNA deamination by constructs ABE8e*A ($k_{obs} = 0.12 \pm 0.01$ min^{-1}) and ABE8e*A (R129E^{cat}) ($k_{obs} = 0.12 \pm 0.01$ min^{-1}) are identical. The DNA deamination rate ($k_{obs} = 0.06 \pm 0.01$ min^{-1}) and efficiency of the ABE8e*A (R129Edock) are significantly reduced (data analysed with one-tailed t-tests, and p = 0.0017). The grey curve represents the single-turnover kinetics of the ABE8E*A mutant containing two amino acid substitutions: one in the docking TadA8e (R129E) and one in Cas9 (E1049R). The introduction of the Cas9 E1049R mutation increases the rate ($k_{obs} = 0.14 \pm 0.01$ min^{-1}) and the efficiency of DNA deamination to the levels of ABE8e*A. For all single-turnover kinetics assays the fraction (in %) of deaminated DNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments.

Based on our computational analysis, R129 in the TadA8e docking unit anchors the E1049 residue in the Cas9 RuvC domain through charge-charge interactions (Figure 6.3d). To probe the role of R129 we adopted a similar approach, where we introduced R129E in the TadA8e catalytic unit [ABE8e*A (R129E^{cat})] and then in the docking domain [ABE8e*A (R129E^{dock})] (Figure 6.4a). In agreement with our computational analysis, R129E in the TadA8e docking domain hampers DNA deamination, while the activity stays unaltered when introduced in the catalytic unit (Figure 6.4c, Supplementary Figure E.10e, f). To further elaborate the role of the interaction between R129 and the Cas9 residue E1049, we prepared a rescue variant with R129E in the TadA8e docking domain and E1049R in Cas9 [ABE8e*A (R129E^{dock}/E1049R^{Cas9})]. The measured deoxyadenosine deamination kinetics showed comparable activity as that of ABE8e*A (Figure 6.4c; Supplementary Figure E.10e). Thus, our rescue substitution could successfully reinstate the ABE8e*A activity, confirming that R129 and E1049 engage in charge-charge interactions and that the TadA8e-RuvC binding is important for efficient DNA deoxyadenosine deamination. Collectively, these findings provide compelling evidence that R98 and R129 residues of the TadA8e docking domain structurally bridge the juxtaposition of TadA8e with Cas9 and the DNA NTS, and are necessary for efficient DNA deamination by ABE8e. As these interactions remarkably weaken in ABE7.10, they uniquely contribute to the enhanced activity of ABE8e.

6.3.3 Key substitutions introduced during ABE8e's evolution tune TadA stability for enhancing DNA deamination capability

TadA8e holds eight amino acid substitutions with respect to TadA7.10.[259,262] This contributes to an astonishingly high 580-fold increase in the rate of deoxyadenosine deamination in ABE8e with respect to ABE7.10, the highest gain across two consecutive generations during the directed evolution. To understand these mutations' role in the complex's stability and shed light on their role in the increased function in ABE8e, we performed free energy simulations. We employed an alchemical free energy perturbation (FEP) method[267] designed to compute the impact of mutations on the folding free energy[268,269] (details in the Methods section and Supplementary Methods), as changes in thermal stability are driven by variations in folding.[268] We computed the relative free energy change in the stability of the ABE7.10 complex while introducing the ABE8e substitutions (i.e., A109S, T111R, D119N, H122N, Y147D, F149Y, T166I, D167N).

Our free energy simulations show that T111R and D119N remarkably alter the stability of the ABE7.10 complex, while the remaining substitutions had moderate effects (Figure 6.5a, Supplementary Figure E.11). The change in free energy of the complex ($\Delta\Delta G$) increases with the T111R mutation, destabilizing ABE7.10, while D119N provides stability to the complex as evidenced from negative $\Delta\Delta G$. To experimentally assess the effect of T111R and D119N on the stability of ABE7.10, we performed a thermal stability analysis using CD spectroscopy and extracted melting temperatures (T_m) for three TadA7.10 heterodimers (composed of wt-TadA and TadA7.10): TadA7.10, TadA7.10 (T111R), and TadA7.10 (T111R/D119N). Notably, the TadA7.10 (T111R)

variant exhibited the lowest thermal stability, with a T_m of 55.4 $\pm$ 0.1 °C, which is 15 °C lower than the T_m of TadA7.10 (70.7 $\pm$ 0.3°C) (Figure 6.5b). However, the introduction of D119N to the TadA7.10 (T111R) variant restored its stability to the TadA7.10 level, elevating the T_m to 69.3 $\pm$ 0.9 °C. This observation corroborates our alchemical FEP simulations, indicating that T111R destabilizes the complex, and contrarily D119N provides stability (Figure 6.5a). This remarkable agreement between our alchemical FEP and thermal stability assays stems from the fact that thermal stability is a key component of the folding free energy computed through our alchemical FEP approach.

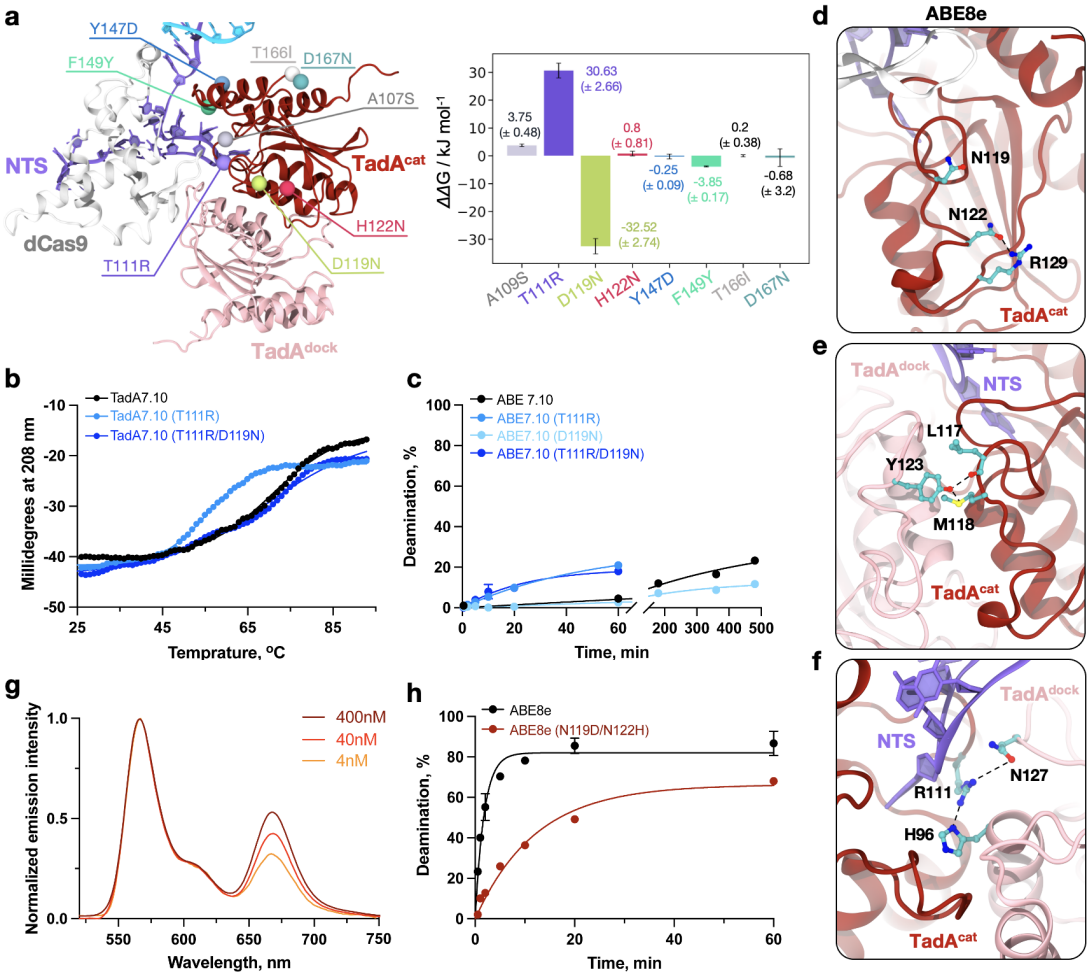

a
Y147D
F149Y
T166I
D167N
A107S
NTS
TadAcat
dCas9
H122N
T111R
D119N
TadAdock
ΔΔG / kJ mol-1
3.75 (± 0.48)
30.63 (± 2.66)
0.8 (± 0.81)
-0.25 (± 0.09)
-3.85 (± 0.17)
0.2 (± 0.38)
-0.68 (± 3.2)
-32.52 (± 2.74)
A109S
T111R
D119N
H122N
Y147D
F149Y
T166I
D167N
b
TadA7.10
TadA7.10 (T111R)
TadA7.10 (T111R/D119N)
Millidegrees at 208 nm
Temprature, °C
c
ABE 7.10
ABE7.10 (T111R)
ABE7.10 (D119N)
ABE7.10 (T111R/D119N)
Deamination, %
Time, min
d
ABE8e
N119
N122
R129
TadAcat
e
TadAdock
NTS
L117
Y123
M118
TadAcat
f
TadAdock
NTS
N127
R111
H96
TadAcat
g
400nM
40nM
4nM
Normalized emission intensity
Wavelength, nm
h
ABE8e
ABE8e (N119D/N122H)
Deamination, %
Time, min

Figure 6.5: **Functional role of amino acid substitutions introduced during directed evolution of ABE8e.** **(a)** Relative free energy changes ($\Delta\Delta G$, in kJ/mol) in the stability of ABE7.10 while introducing the ABE8e mutations, computed through alchemical free energy perturbation (FEP) simulations. The amino acid substitutions considered for alchemical FEP are shown on the three-dimensional structure of ABE8e on the right. **(b)** Thermal denaturation profiles of TadA7.10 variants [TadA7.10 (black), TadA7.10 (T111R) (blue) and TadA7.10 (T111R/D119N) (dark blue)] were measured by CD spectroscopy with a temperature change of 1 °C/min. The thermal denaturation data were fitted to the Boltzmann sigmoidal equation and the T_m values for each TadA7.10 variant were determined by the inflection point and are 70.7 $\pm$ 0.3°C, 55.4 $\pm$ 0.1°C, and 69.3 $\pm$ 0.9°C, respectively. **(c)** Single-turnover kinetics of ABE7.10 (black), ABE7.10 (T111R) (blue), ABE7.10 (D119N) (light blue), and ABE7.10 (T111R/D119N) (dark blue) measured using fluorescently labelled dsDNA containing a single adenine within the editing window. The observed rates of DNA deamination by these constructs are 0.0021 $\pm$ 0.0005 min^{-1}, 0.020 $\pm$ 0.003 min^{-1}, 0.0038 $\pm$ 0.0007 min^{-1}, 0.020 $\pm$ 0.003 min^{-1}, 0.042 $\pm$ 0.007 min^{-1}, respectively. The fraction (in %) of deaminated DNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments. **(d-f)** Interactions established by the solvent-exposed loop of the TadA catalytic domain (residues 108-125) in ABE8e. N122 forms a stable interaction with R129 within the same loop (d). Y123 of the TadA8e docking domain interacts with L117 and M118 of the TadA catalytic domain (e). R111 of the TadA8e catalytic domain interacts with the H96 and N127 residues of the docking domain (f). Detailed analysis from MD simulations and comparison with ABE7.10 are reported in Supplementary Figures E.14-E.15. **(g)** Fluorescence emission spectra of Cy3/Cy5-labeled ABE8e (N119D/N122H) across varying concentrations, with normalized Cy3 emission intensity, highlighting the changes in FRET signal. The observed distinct reduction in Cy5 fluorescence emission intensity (at 675 nm), as protein concentration decreases, clearly indicates that N119D and N122H substitutions destabilize TadA8e dimer. **(h)** Single-turnover kinetics of ABE8e (black) and ABE8e (N119D/N122H) (red) measured using fluorescently labelled dsDNA containing a single adenine within the editing window. The observed rates of DNA deamination by these constructs are 0.56 $\pm$ 0.06 min^{-1} and 0.08 $\pm$ 0.01 min^{-1}, respectively. The fraction (in %) of deaminated DNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments.

Interestingly, these observations align with several studies performed to understand the role of single amino acid substitutions that appear during directed evolution. Independent studies by Romero & Arnold,[270] Tawfik,[271] and Shoichet[272] demonstrated that destabilizing substitutions are accommodated during directed evolution to reshape functionality (i.e., providing improved or new function), while structural, stabilizing mutations to compensate for the loss of stability. We thereby monitored the single-turnover deamination rate of ABE7.10 and its mutants, ABE7.10 (T111R), ABE7.10 (D119N), and ABE7.10 (T111R/D119N) (Figure 6.5c, Supplementary Figure E.12). Introducing T111R increases the deamination rate by approximately 10-fold with respect to ABE7.10, indicating the functional role of this destabilizing mutation. D119N by itself has a slight negative effect on deamination, while the deamination rate of the ABE7.10 (T111R/D119N) is comparable to the rate of ABE7.10 (T111R), indicating that D119N does not contribute to the enhancement of DNA deamination rate, but rather stabilizes the ABE7.10. Hence, the destabilizing T111R substitution has a functional role, while D119N compensates for the loss of stability during evolution. The above-mentioned studies on directed evolution also indicated that mutations providing improved (or new) catalytic function often appear first and destabilize the complex. In the evolution of ABE8e from ABE7.10, the T111R mutation frequently appeared in several early variants,[262] enhancing the ABE7.10 deamination rate according to our single-turnover kinetics. Since T111R reduces complex stability, the subsequent introduction of D119N mutation compensates for this, facilitating the enzyme's evolution. Together, T111R and D119N introduce features that balance activity and stability, driving the evolution from ABE7.10 to the improved ABE8e.

To further understand the biophysical role of T111R and D119N in the activity of ABE8e, we analysed our multi-μs long simulations of ABE7.10 and ABE8e. Notably, the T111R and D119N substitutions reside in a solvent-exposed loop of the TadA8e catalytic domain (residues 108-125). This loop exhibits significantly reduced flexibility in ABE8e when compared to ABE7.10 (Supplementary Figure E.13). This loss in flexibility is attributed to altered interactions involving D119N and a few other mutations. The backbone interaction between D119 and H122 in ABE7.10 is remarkably weakened in ABE8e and a new persistent interaction is established between N122 and R129 in ABE8e (Figure 6.5d, Supplementary Figure E.14). We then evaluated whether the flexibility of this loop could alter the interactions between the TadA monomers (Supplementary Figure E.15). We observe that H123Y in the TadA8e docking domain improves interaction with the L117 backbone (Figure 6.5e), with respect to H123 in TadA7.10. Additionally, the reduced flexibility of the loop reinforces the interactions between T111R of the catalytic domain and the H96 and N127 residues of the docking domain (Figure 6.5f), which is also evidenced by a gain in the strength of these interactions in ABE8e vs. ABE7.10 (Figure 6.3b). This promotes the engagement between the catalytic and the docking TadA8e units.

To experimentally assess the role of D119N and N122H in the dimerization of TadA, we performed ensemble FRET experiments. We generated FRET constructs of the ABE8e containing N119D and N122H substitutions as seen in its predecessors TadA7.10 as well as wt-TadA. The Cy3/Cy5-labelled ABE8e (N119D/N122H) is properly folded (Supplementary Figure E.16a), and the dye-labelling does not affect the

DNA deamination rate (Supplementary Figure E.16b,c). Our ensemble FRET experiment shows that TadA8e (N119D/N122H) enzymes exist in dimeric form when fused to dCas9, as evidenced by the presence of a peak at ∼675 nm in the fluorescence emission spectra of Cy3/Cy5-labelled proteins when excited at 530 nm. However, the dimerization affinity of this TadA8e variant is lower compared to that of TadA8e. It appears to be similar to the affinity of the wt-TadA, as indicated by the reduction in FRET signal with decreasing protein concentrations (Figure 6.2c, 6.5g). Since N119D and N122H destabilise the TadA8e dimer, based on our observation that TadA dimerization is important for the enhanced activity of the ABE8e (Figure 6.2), we anticipate that the ABE8e variant with these substitutions – although they are distant from the catalytic site – will exhibit lower activity compared to ABE8e. Indeed, our single turnover kinetic assays show that the DNA deamination rate of ABE8e (N119D/N122H) (k_{obs} = 0.08 ± 0.01 min^{-1}) is seven times lower than the rate of ABE8e (k_{obs} = 0.56 ± 0.06 min^{-1}) (Figure 6.5h). Collectively, experimental and computational observations demonstrate that the D119N and H122N, introduced to the ABE8e during directed evolution reduces the flexibility of the loop and stabilize inter-molecular interactions. These findings agree with the previous observations demonstrating the importance of this loop in DNA deamination activity of first generations of TadA.[273] Moreover, it aligns with the growing recognition of the importance of loop dynamics in protein evolution, with several studies indicating that loop motion can influence catalysis.[274,275]

6.4 Discussion

Adenine base editors (ABEs) hold tremendous potential in treating genetic diseases caused by point mutations. They harness the precision of CRISPR-Cas systems and the catalytic power of laboratory-evolved TadA enzymes to facilitate the conversion of A•T to G•C via the deamination of deoxyadenosine to deoxyinosine at specific ss-DNA sites.[253] Among the ABEs used for precision genome editing in various cells and organisms, ABE8e is one of the most efficient DNA base editors to date.[262,263] It demonstrates a notable leap forward compared to ABE7.10, its closest precursor, dramatically enhancing both DNA base editing efficiency and the rate of DNA deamination. However, ABE8e is more prone to DNA off-target editing.[262] A thorough understanding of this improvement is instrumental in advancing the development of efficient ABEs. Here, we employed an integrative approach where we combined extensive MD simulations with free energy methods, and biochemical and biophysical experiments to establish the biophysical rationale behind the astonishing base editing efficiency of ABE8e, providing new directions for the development of more advanced ABEs.

ABEs have been developed to deaminate DNA by evolving the wt-TadA,[259,262] a natural homo-dimeric enzyme that catalyses the deamination of tRNA with its active site formed by both monomers. The cryo-EM structure of ABE8e in a DNA-bound state reveals that TadA8e forms a similar dimer (Figure 6.1). This raises the question of how amino acid substitutions introduced during directed evolution impact TadAs' dimerization and how the latter affects DNA base editing. Using enhanced sampling simulations, viz., an Umbrella Sampling method, in combination with FRET measurements, we show

that TadA8e forms more stable dimers than its progenitor wt-TadA (Figure 6.2). Interestingly, during the directed evolution of the wt-TadA deaminase toward TadA8e, the selective pressure was applied to select for efficient DNA base editing rather than for more stable dimers. This suggests that TadA dimerization strength co-evolved with the DNA base-editing property, highlighting the critical role of TadA dimerization in the DNA deamination activity.

In light of these findings, we sought to establish what makes the dimeric TadA8e an efficient DNA base editor when bound to the dCas9-DNA complex. Extensive MD simulations of ABE8e, in comparison with ABE7.10, its closest predecessor with ∼580-fold lower DNA deamination rate, and ABE0.1 (where TadA is wt) provide evidence that the TadA8e dimer establishes key interactions with both the DNA and dCas9, a phenomenon uniquely observed in ABE8e (Figure 6.3). Indeed, TadA8e forms a robust network of interactions involving its docking domain, the two antiparallel β-sheets of RuvC (residues 1049 to 1056), and the DNA NTS, consolidating the positioning of TadA8e within the dCas9-DNA complex. Specific residues within the TadA8e docking domain, R98 and R129, are found to form essential interactions with the DNA NTS and the two antiparallel β-sheets of RuvC (Figure 6.3). Single-turnover kinetics demonstrated that these interactions are necessary for efficient DNA deamination by ABE8e (Figure 6.4) and, considering that they are remarkably weaker in ABE7.10 (Figure 6.3), they critically contribute to the activity of ABE8e. This discovery holds significant implications for developing of new ABEs, particularly when employing orthogonal Cas9 or other Cas effectors, in which the antiparallel β-sheet of RuvC that binds TadA8e is absent or is chemically and structurally different. It highlights the

necessity of engineering the Cas9-docking TadA interface for each distinct Cas effector to achieve optimal interactions and, in turn, efficient DNA base editing. This aligns with prior observations that fusing TadA8e to different Cas effectors results in less efficient DNA base editing when compared to the ABE8e (TadA8e fusion to the Cas9 from *S. pyogenes*).[262]

Reinforcing the interactions between TadA8e and the Cas effector could enable them to synergize more effectively, potentially also leading to a reduction in DNA off-target editing. Indeed, the fundamental process driving ABE8e's bystander editing is proposed to stem from TadA8e being a multiple-turnover enzyme and its flexible connection to Cas9 enabling it to modify nearby ssDNA regardless of whether Cas9 is bound to it or not.[258,259,262] Hence, enhancing the interactions between TadA8e and the Cas effector could serve as a strategy to reduce ABE8e's bystander editing and Cas9-independent DNA off-target effects.

ABE8e holds eight amino acid substations in its TadA8e unit with respect to its closest predecessor ABE7.10, contributing somehow to enhancing ABE8e's superior efficiency and rate of DNA deamination compared to ABE7.10. Our alchemical free energy perturbation simulations and thermal stability assays reveal that two substitutions, T111R and D119N, are key for the TadA8e stability and function. We show that T111R is a destabilizing mutation that provides TadA with an improved catalytic function (Figure 6.5). In contrast, D119N is a compensatory, stabilizing mutation and does not directly enhance TadA's activity. This dynamic interplay between destabilizing and stabilizing mutations aligns with the complex nature of enzyme evolution and adaptation. Biophysical studies of evolved enzymes have shown that mutations conferring improved

or new catalytic function commonly appear early and destabilize the complex.[270–272] As the adaptive process continues, proteins regain stability through other mutations with little to no impact on activity. Accordingly, moving from ABE7.10 to ABE8e, T111R appeared in most of the evolution batches, followed by D119N.[262] This suggests that the interplay between T111R and D119N provides the balance between activity and stability necessary to evolve the complex toward improved function. This underscores the importance of employing diverse approaches when engineering enzymes, as mutations impacting the complex stability and activity are not readily identifiable through rational design alone.

In addition to TadA stabilization, the D119N in combination with H122N substitution plays a crucial role in stabilizing the TadA8e dimer by altering the dynamics of the 105-125 loop region, as evidenced by our MD simulations and ensemble FRET experiments. Since a stable TadA8e dimer is essential for forming interactions with Cas9 and DNA,[276,277] reversing these substitutions destabilizes the TadA8e dimer, subsequently reducing the activity of ABE8e (Figure 6.5). This is in line with the significance of loop dynamics in protein evolution, as numerous studies suggest that loop motion directly or indirectly affect catalysis.[274,275] Therefore, we propose that the 105-125 loop is an important region that could be further engineered to develop more efficient ABEs.

In summary, extensive molecular simulations, single turnover kinetics and biophysical experiments reveal that TadA dimerization and its juxtaposition to the dCas9-DNA complex are pivotal for the efficient DNA deamination in ABE8e. This finding is critical for engineering superior ABEs utilizing orthogonal Cas9 or other Cas effectors to reduce the editor size improving its delivery to cells, and to potentially reduce DNA off-target

editing, thereby contributing to the advancement of safer and more efficient genome editing tools.

6.5 Materials & Methods

6.5.1 Structural Models

Molecular simulations were based on the cryo-EM structure of the ABE8e complex (PDB 6VPC)[259] at 3.20 Å resolution, comprising a catalytically dead Cas9 (dCas9) from *S. pyogenes* and the TadA8e enzyme. Using this structure, which preserves the bona fide configuration of the TadA enzymes bound dCas9, four model systems were built: ABE0.1 (where TadA is WT), ABE7.10, ABE8e, and an ABE8e including the TadA8e heterodimer, in which the catalytic domain is the evolved TadA8e and the docking domain is the wt-TadA (referred to as ABE8e hetero). The X-ray structure of wt-TadA (PDB 1Z3A,[261] 2.03 Å resolution) was used to model the wt-TadA and heterodimer docking domain. The TadA7.10 catalytic domain was modelled by mutating the ABE8e structure (PDB 6VPC). Missing residues were added via AlphaFold2.[278] All systems were solvated, neutralized, and resulted in periodic cells ($\sim$133.68 × 148.43 × 224.94 $Å^3$) with $\sim$500,000 atoms each. Details are in the Supplementary Methods.

6.5.2 Molecular Dynamics (MD) Simulations

MD simulations were performed using a simulation protocol tailored for RNA/DNA nucleases,[279] which we also used in studies of CRISPR-Cas systems,[56,168,280] embracing the use of the Amber ff19SB[126] force field, and the OL15[281] and OL3[102] corrections for

DNA and RNA respectively. The TIP3P[104] model was used for water molecules. We re-parametrized the Zn^{2+} ion and the active site residues using Quantum Mechanical (B3LYP[282]/6-31G*[283]) calculations, following the protocol by Li & Merz[284,285] Upon energy minimization, the systems were heated from 0 to 100 K in two consecutive NVT simulations (representing the canonical ensemble) of 5 ps each, imposing positional restraints of 100 kcal mol^{-1} $Å^{-2}$ on the protein, the nucleic acids, and the TadA active site. Further heating to 200 K occurred in a 100 ps run in the isothermal-isobaric ensemble (NPT), reducing restraints to 25 mol^{-1} $Å^{-2}$. Restraints were removed, heating to 300 K in a 500 ps NPT simulation. These simulations were performed using a 1 fs time step. The time step was subsequently increased to 2 fs for further equilibration and production runs. Hydrogen bonds were constrained with the SHAKE algorithm.[138] After ∼600 ps of equilibration, ∼10 ns NPT simulations stabilized the systems' density around 1.01 g/cm^3. The temperature was kept constant at 300 K via Langevin dynamics, with a collision frequency $\gamma = 1/ps$. The pressure was controlled in the NPT simulations by coupling the system to a Berendsen barostat[108] at a reference pressure of 1 atm and with a relaxation time of 2 ps. Subsequently, production runs were carried out in the NVT ensemble, reaching ∼3 μs for each system and in triplicate (i.e., ∼9 μs of aggregate sampling per system, totalling ∼36 μs across four systems). MD simulations replicates were performed starting from different configurations and velocities initialized according to the Maxwell-Boltzmann distribution at physiological temperature. These simulations were performed using the GPU-empowered version of AMBER 22.[131] Details are reported in the Supplementary Methods.

6.5.3 Umbrella Sampling Simulations

The Umbrella Sampling (US)[117] method was used to compute the free energy profiles associated with the dimerization of TadA enzymes. In this method, several simulations (US windows) are run in parallel with additional harmonic bias potential applied to selected Reaction Coordinates (*RC*):

$$V(RC) = \frac{k}{2}(RC(t) - RC^*)^2 \tag{6.1}$$

where $V(RC)$ is the value of the bias potential, k is a bias force constant, $RC(t)$ is the value of *RC* at given time t and $RC*$ is the reference value of RC. By using different $RC*$ values in each US window, one can sample the biased probability distribution $p_b(RC)$ along the whole *RC* range of interest. Thanks to a reweighting scheme providing unbiased probabilities, the free energy of a process along the RCs can be computed (*vide infra*).

We performed US simulations at 300 K using two RCs. RC1 was the inter-monomeric distance considering the center of mass of each TadA unit. RC2 was the difference in root-mean-square deviation (RMSD) of all the Cα atoms with respect to the experimental structures (i.e., PDB 1Z3A[261] for wt-TadA and PDB 6VPC[259] for TadA8e). The combination of these RCs enabled us to describe the dimerization process while keeping the experimental structures as a reference. Restraint forces were applied on the two RCs as follows. RC1 was discretized into bin sizes of 0.5 Å with a collective force constant k of 17.5 kcal mol^{-1} Å^{-2}. RC2 was divided into bin sizes of 0.5 Å RMSD difference with a collective force constant k of 25 kcal mol^{-1} Å^{-2}. Each window was run for

15 ns. Overall, we simulated 51 (RC1) * 31 (RC2) = 1,581 windows, resulting in a collective ensemble of $\sim$25 μs of sampling for each system. Two independent sets of two-dimensional (2D) US simulations were performed considering the wt-TadA and TadA8e. The free energy profiles were computed using the Variational Free Energy Profile (VFEP) method,[286] which uses the maximum likelihood principle to construct free energy profiles. This method was applied upon removing the initial $1/3^{rd}$ of the US trajectories from each window, considering the system's relaxation. Analysis of the minimum free energy pathway (MFEP) between the dimeric and monomeric states was performed using the MEPSA code,[140] which employs an approach similar to the Dijkstra algorithm to trace the MFEP between nodes of interest. The convergence of the free energy profiles was evaluated by plotting the minimum free energy pathways, considering the remaining $\sim$10 ns runs from each US window, for $\sim$5 ns through $\sim$10 ns (Supplementary Figure E.3). The error estimation on the minimum free energy pathways was performed using the Monte Carlo bootstrap error analysis.

6.5.4 Alchemical Free Energy Perturbation

We performed alchemical free energy perturbation (FEP) simulations using lambda (λ) dynamics,[267] to compute the relative free energy change (i.e., $\Delta\Delta G$) in the ABE7.10 complex while introducing the ABE8e mutations (i.e., A109S, T111R, D119N, H122N, Y147D, F149Y, T166I, D167N). With our alchemical FEP simulations, we aimed to describe the effect of amino acid mutations on the thermal stability of the complex. Therefore, we utilized an alchemical FEP method to calculate the effect of mutations on the folding free energy,[268,269] since changes in thermal stability result from modifications

in the free energy of folding. Consistently with a thermodynamic cycle (details in the Supplementary Methods), state *A* was evolved into state *B*, representing the initial and the final residues respectively, via a series of transition states following the alchemical free energy formalism. A lambda (λ) dynamic variable was used to define the transitioning thermodynamic states of the system, where $\lambda = 0$ and $\lambda = 1$ represent the initial and final states respectively. The distribution of the λ parameter was optimized using a gradient descent algorithm to maximize the probabilities of exchange between adjacent states (https://gitlab.com/KomBioMol/converge_lambdas).[287] In detail, the neighbouring λ-windows were allowed to exchange their configurations every 0.5 ps according to the Metropolis criterion, and the values of λ were optimized to achieve the acceptance rate of at least 15 %. Since each intermediate λ state is technically a hybrid between the *A* and *B* end point states, we generated their dual coordinates and topologies using the PMX server (http://pmx.mpibpc.mpg.de/).[288] The relative free energy changes (i.e., $\Delta\Delta G$) for the transformation were computed using the Multistate Bennett Acceptance Ratio (MBAR)[118,119] method to integrate the free energies over the different λ values.[143,144] Using this method, we computed relative $\Delta\Delta G$ for each single-point mutation introduced through directed evolution in ABE7.10 leading to ABE8e (i.e., A109S, T111R, D119N, H122N, Y147D, F149Y, T166I, D167N). Alchemical FEP simulations were performed across the lambda values for each mutation, running $\sim$50 ns for each window and totalling $\sim$8 μs sampling. The convergence of the $\Delta\Delta G$ for each system is shown in Supplementary Figure E.11.

6.5.5 Preparation of ABE-sgRNA Samples

ABE proteins were expressed and purified as described previously by Lapinaite et al. Single guide RNA (sgRNA) was transcribed overnight at 37 °C using in-house prepared T7 polymerase and purified by using 7 M urea, 10 % 19:1 polyacrylamide gel electrophoresis, followed by crush-and-soak extraction overnight at 4 °C and ethanol precipitation as described previously.[259] The ABE-sgRNA (RNP) complexes were formed by incubating the protein (Supplementary Table E.1) with the refolded sgRNA at a molar ratio of 1:1.5 (protein to sgRNA) in RNP assembly buffer (20 mM Tris-HCl, 200 mM KCl, 5 mM $MgCl_2$, 5% (v/v) glycerol, 2 mM DTT, pH 7.5) at room temperature for 20 minutes. More details are provided in the Supplementary Methods.

6.5.6 Single-turnover *in vitro* DNA and RNA Deamination Assays

The single turnover DNA deamination reactions were conducted with 2 μM ABE RNPs in a deamination buffer consisting of 20 mM Tris-HCl (pH 7.5), 100 mM KCl, 5 % (v/v) glycerol, 2.5 mM $MgSO_4$, and 2 mM DTT. These reactions were initiated by introducing fluorescently labelled dsDNA or hairpin RNA resembling anticodon hairpin of $tRNA^{Arg2}$ to achieve a final concentration of 7.5 nM. The reactions were incubated at 37°C. 20 μ, aliquots were transferred into 30 μL preheated water at 95 °C at different time intervals (0, 1 min, 2 min, 5 min, 10 min, 20 min and 60 min for ABE8e and ABE0.1 variants and 0, 1 min, 2 min, 5 min, 10 min, 20 min, 60 min, 180 min, 360 min, and 480 min for ABE 7.10 variants) and quenched by heating at 95 °C for 5 min. For detection of DNA deamination the quenched and purified reactions were incubated with 20 units of EndoV (NEB) at 37 °C for 1.5 hours. For detection

of RNA deamination, human EndoV (hEndoV; purified in-house) was added into the reaction at a final concentration of 0.1 μM. The reactions were incubated at 37 °C for 1 h. Samples were resolved on 15 % (v/v) denaturing (29:1) polyacrylamide gel. The gel was visualized by using the fluorescence imaging instrument Amersham Typhoon (GE Healthcare). The experiments were conducted in three separate replicates, and the levels of uncleaved and cleaved DNA or RNA were evaluated using ImageQuant TL Software from GE Healthcare. More details are provided in the Supplementary Methods.

6.5.7 Thermal Stability Assay using CD Spectroscopy

The sample was prepared and loaded into the 1 mm pathlength microcuvette (Hellma 110-1-40, style 110-QS). The measurements were performed using the Jasco J-815 spectropolarimeter. The ellipticity at the wavelength of 208 nm was recorded from 20 °C to 95 °C at the ramp rate of 1 °C/min. The measurement was performed with an integration time of 2 seconds, using a bandwidth of 1 nm. The ellipticity was plotted as a function of temperature and fitted with a Boltzmann sigmoidal equation. The melting temperature, T_m, was determined by the inflection point.

6.5.8 Fluorescence Resonance Energy Transfer (FRET) Measurements

Briefly, the labelling reactions of ABEs were carried out in SEC buffer and contained 10 μM protein along with 200 μM Cy3- and 1.2 mM Cy5-maleimide (APExBIO) (1:6 ratio of donor to acceptor). The labelling reaction was initiated by gentle rocking in the dark for 2 hours at room temperature, followed by overnight incubation at 4 °C.

Reactions were quenched by adding 10 mM DTT and unreacted dyes were removed via SEC. The purified labelled ABE samples were concentrated, flash-frozen in liquid nitrogen, and stored at -80 °C. All fluorescence measurements were conducted at room temperature in SEC buffer. Fluorescence measurements were performed on ABE-sgRNA samples containing various concentrations (400 nM, 40 nM and 4 nM) of Cy3/Cy5-labelled ABE0.1 [wt-TadA(T133C/C141S)-dCas9(C80S/C574S)] or ABE8e [TadA8e(T133C/C141S)-dCas9(C80S/C574S)]. Fluorescence measurements were performed using a 10 mm path-length quartz cuvette (Hellma Analytics), a QM-4/2005SE spectrofluorometer and a PTI 814 Photomultiplier, using 5 nm slit widths and 0.5 second integration time. For each sample, the fluorescence emission spectra were collected, in which the sample was excited at 500 nm and emission was recorded from 510 to 750 nm in 1 nm increments. Data was collected and processed using PTI's FeliX32 software. Experiments were performed at least three times. The FRET spectra obtained on samples at decreasing concentrations were normalized to the maximum Cy3 donor peak at 570 nm to eliminate the effect of sample dilution. Spectra were smoothed using the Savistsky-Golay method and plotted using Prism (GraphPad Software). More details are provided in the Supplementary Methods.

6.6 Acknowledgements

We thank Dr. Marcia Levitus for her helpful discussions in setting up ensemble FRET measurements and providing access to the spectrophotometer. This material is based upon work supported by the National Institutes of Health (Grant No. R01GM141329, to

G.P. and DP2GM149550 to A.L.) and the National Science Foundation (Grant No. CHE-2144823, to G.P.). GP acknowledges support by the Sloan Foundation (Grant No. FG-2023-20431) and the Camille and Henry Dreyfus Foundation (Grant No. TC-24-063). This work used Expanse at the San Diego Supercomputing Center through allocation MCB160059 and Bridges2 at the Pittsburgh Supercomputer Center through allocation BIO230007 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation supports grants #2138259, #2138286, #2138307, #2137603, and #2138296. Computer time was also provided by the National Energy Research Scientific Computing Center (NERSC) under Grant No M3807.

Appendix A

Supporting Information for Chapter II

A.1 Structural models

Molecular simulations have been performed on five model systems of CRISPR-Cas12a, based on the available crystallographic structures. The RNA-bound states are based on two X-ray structures: *Lachnospiraceae bacterium* Cas12a (LbCas12a) solved at 2.38 Å resolution (5ID6.pdb)[43] and *Francisella novicida* Cas12a (FnCas12a) solved at 3.34 Å resolution (5NG6.pdb).[2] The DNA-bound states have been based on three X-ray structures. We considered the structures of *Acidaminococcus sp.* Cas12a (AsCas12a) and FnCas12a, in which the NTS is partially cleaved (i.e., the 5B43.pdb at 2.38 Å resolution[41] and the 5NFV.pdb at 2.50 Å resolution,[2] respectively). The third DNA-bound state has been based on a recent structure of FnCas12a, including a longer NTS that binds within the RuvC cleft and reconciles with the TS (6I1K.pdb at 2.65 Å resolution).[16] Missing residues of the X-ray structures have been added via homology modeling, using SwissModel by Schwede.[289] Missing bases of 6I1K X-ray structure

have been modeled by Jinek in the original paper.[16] The simulation systems have been embedded in explicit waters, while Na+ ions were added to neutralize the total charge, leading to orthorhombic periodic simulation cells of 145 × 135 × 115 $Å^3$ (RNA-bound LbCas12a, for a total of ∼200,000 atoms), ∼115 × 130 × 150 $Å^3$ (RNA-bound FnCas12a, ∼200,000 atoms), ∼125 × 125 × 150 $Å^3$ (DNA-bound AcCas12a, ∼210,000 atoms), ∼120 × 135 × 150 $Å^3$ (DNA-bound FnCas12a, ∼210,000 atoms) and ∼120 × 135 × 150 $Å^3$ (DNA-bound FnCas12a', ∼220,000 atoms).

A.2 Molecular Dynamics (MD)

The above-mentioned model systems have been object of MD, adopting a well-established simulation protocol for protein/nucleic acid complexes.[290] We employed Amber ff12SB force field, which includes the ff99bsc0[100] corrections for DNA and the ff99bsc0+χOL3[101,102] corrections for RNA. The Allnér force field[103] has been employed for $Mg^{2}+$ ions and TIP3P model[104] for waters. These force field parameters and the simulation protocol have also been employed in our recent studies of CRISPR-Cas9,[73–75,88,94] enabling a fair comparison. An integration time step of 2 fs has been employed. All bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm. Temperature control (300 K) has been performed via Langevin dynamics,[107] with a collision frequency γ = 1/ps. Pressure control was accomplished by coupling the system to a Berendsen barostat,[108] at a reference pressure of 1 atm and with a relaxation time of 2 ps. The systems have been subjected to energy minimization to relax water molecules and counter ions, keeping the protein, the RNA, DNA and Mg^{2+}

ions fixed with harmonic position restraints of 300 kcal/mol $\cdot$ Å^2. Then, the systems have been heated up from 0 to 100 K in the canonical ensemble (NVT), by running two simulations of 5 ps each, imposing position restraints of 100 kcal/mol $\cdot$ Å^2 on the above-mentioned elements of each system. The temperature was further increased up to 200 K in $\sim$100 ps of MD in the isothermal-isobaric ensemble (NPT), reducing the restraint to 25 kcal/mol Å^2. Subsequently, all restraints were released and the temperature of the systems was raised up to 300 K in a single NPT simulation of 500 ps. After $\sim$1.1 ns of equilibration, $\sim$10 ns of NPT runs were carried out allowing the density of the system to stabilize around 1.01 g/cm^{-3}. Finally, $\sim$1 μs of MD simulations was carried out in NVT ensemble for each system, which has also been simulated in four independent replicates. The MD simulations replicas have been obtained starting from different configurations and velocities, initialized accordingly to the Maxwell-Boltzmann distribution at physiological temperature. Considering that we considered five simulation systems, we collected a total of $\sim$20 μs of aggregate sampling (i.e., 5 systems * 4 replicas * $\sim$1 μs = $\sim$20 μs). These simulations have been performed using the GPU version of AMBER 18.[109]

A.3 Principal Component Analysis (PCA)

PCA is a statistical method that can report on the large-scale, collective motions occurring in biological macromolecules undergoing MD simulations. Thus, PCA can provide valuable information on major conformational changes taking place along the trajectory. In fact, through this statistical technique, it is possible to reduce the large number of

degrees of freedom to an essential subspace set, which captures large-amplitude motions of the system. In PCA, the covariance matrix of the protein Cα atoms is calculated and diagonalized to obtain a new set of coordinates (eigenvectors) to describe the system motions. Each eigenvector – also called Principal Component (PC) – is associated to an eigenvalue corresponding to the mean square fluctuation contained in the system's trajectory projected along that eigenvector. By sorting the eigenvectors according to their eigenvalues, the first PC (i.e., PC1) corresponds to the system's largest amplitude motion, and the dynamics of the system along PC1 is usually referred as "essential dynamics".[87,291]

Here, the principal motions of the protein have been captured starting from the mass-weighted covariance matrix of the Cα atoms. In detail, PCA has been performed considering the FnCas12a systems, whereby the collected ensembles (i.e., arising from the compared RNA-bound FnCas12a, DNA-bound FnCas12a and FnCas12a' systems; Figure 2.3) were combined and subjected to RMS-fit to the same reference configuration, removing the rotational and translational motions. This was performed to ensure a consistent eigenbasis and motions of the PCs on all compared systems and to construct the covariance matrices from the atoms' positions. We also performed two independent PCA on the LbCas12a and AsCas12a systems, providing insights into their essential motions. Each element in the covariance matrix is the covariance between atoms i and j, defining the i, j position of the matrix. The covariance C_{ij} is defined as:

$$C_{ij} = \langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle \tag{1.1}$$

where $\vec{r}_i$ and $\vec{r}_i$ are the position vectors of atoms i and j, and the brackets denote an average over the sampled time period. A positive sign of this product indicates that the two atoms move in a correlated manner, otherwise, a negative value points to anti-correlated atoms. If the product is zero, then it evinces that the atoms displacements are independent of each other. The covariance matrix was then diagonalized, leading to a complete set of orthogonal collective eigenvectors, each associated to a corresponding eigenvalue. The eigenvalues denote how much each eigenvector is representative of the system dynamics, thus giving a measure of the contribution of each eigenvector to the total variance. Indeed, the eigenvectors with the largest eigenvalues correspond to the most relevant motions. By projecting the displacements vectors of each atom along the trajectory onto the eigenvectors (i.e., by taking the dot product between the two vectors at each frame), the Principal Components (PC) were then obtained. The cumulative variance accounted by all the PCs was calculated for all system (Supplementary Figures A.5), revealing that the first PCs account for the major contribution. The first PC (i.e., PC1), commonly referred to as "essential dynamics",[87,291] is plotted on the 3D structures of the CRISPR-Cas12a complex indicating the largest amplitude motions of the protein (Figure 2.3a). Upon projecting each structure arising from the MD simulations into the collective coordinate space derived by the first two eigenvectors (PC1 and PC2), plotting PC1 against PC2 allows generation of scatter plots (PC1 *vs.* PC2) that display how the conformational space defined by the first two modes is sampled through the MD simulations (Figure 2.3b).

For each simulated system of Cas12a, PCA has been performed considering the aggregate sampling arising from multiple replicas of $\sim$1 μs each. Indeed, a single MD

run is not sufficient to explore the conformational ensemble of large ribonucleoprotein complexes.[73,93,292] On the other hand, multiple runs starting from different coordinates and initial velocities can more properly explore the conformational states and provide a better description of the protein conformational ensemble through PCA. Accordingly, PCA analysis over the aggregate multi-μs runs have shown to properly depict the conformational ensemble of CRISPR-Cas9 in our previous study.[73] PCA has been performed using cpptraj of Amber18,[109] while the Normal Mode Wizard[293] plugin of the Visual Molecular Dynamics[294] program has been used for the graphical rendering.

A.4 Volumetric Analysis

To calculate the volume of the groove between the REC and NUC lobes, we performed volumetric analysis by employing the POcket Volume MEasurer (POVME) software (v. 3.0).[295] The volumetric analysis has been performed on the collective sampling arising from the equilibrium MD trajectories of the DNA-bound AsCas12a, FnCas12a and FnCas12a'. To calculate the volume of the groove for a given trajectory frame, we monitored the space within two 20 Å radius spheres, which defined the groove between the REC and NUC lobes. In detail, the two spheres have been centered at the center of mass between the Nuc and REC2 domains and at the center of mass between the REC, WED and RuvC respectively (Supplementary Figure A.6). The pocket volume was calculated from the portion of the sphere that was unoccupied by protein acid atoms. To provide a comparable measure of the change in volume and maintain consistency, nucleic acids have been omitted from volumetric analysis. Only volumetric regions

contiguous with the pocket were included. Grid spacing and padding parameters were set to 1.0 and 1.09 Å respectively.

A.5 Cross-Correlation Analysis

Cross-Correlation (CC_{ij}) analysis has been performed in order to identify the dynamical coupling of the motions between Cα atoms (i and j) in the simulated systems. CC_{ij} analysis is based on Pearson coefficients, which are computed between the fluctuations of the Cα atoms relative to their average positions, according to:

$$CC_{ij} = \frac{\langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle}{[(\langle \vec{r}_i^{\,2} \rangle - \langle \vec{r}_i \rangle^2)(\langle \vec{r}_j^{\,2} \rangle - \langle \vec{r}_j \rangle^2)]} \tag{1.2}$$

where $\vec{r}_i$ and $\vec{r}_j$ are the position vectors of atoms i and j, considered over the sampled time period (denoted using brackets). CC_{ij} arise from the normalization of the covariance between atoms i and j (CC_{ij}) reported in Equation 1.1. Positive values of the CC_{ij} coefficients indicate lockstep motions between atoms i and j; while negative CC_{ij} values are indicative of anticorrelated motions. The magnitude of CC_{ij} coefficients (i.e., ranging from 0 – 1 for lockstep motions; and from –1 – 0 for anticorrelated motions) indicates strength of the correlation. In this work, the CC_{ij} coefficients have been plotted as a two–by–two matrix, describing the correlated motions of each residue (i.e., their Cα atoms) with each other. For each model system, the CC_{ij} matrices have been computed over independent $\sim$1 μs trajectories, and have been subsequently averaged over the aggregate sampling arising from 4 independent MD replicas. The CC_{ij} matrices

computed over the individual trajectories and the averaged CC_{ij} matrices are reported in Supplementary Figures A.8-A.12.

A.6 Generalized Correlation (GC_{ij}) Analysis based on Mutual Information

The Generalized Correlations (GC_{ij}) analysis describes correlations independently on the relative orientation of the atomic fluctuations and is able to capture non–linear correlations.[92] In information theory, two variables, such as the such as the $\vec{r}_i$ and $\vec{r}_j$ position vectors, can be considered correlated when their joint probability distribution, $p(\vec{r}_i,\vec{r}_j)$, is smaller than the product of their marginal distributions, $p(\vec{r}_i)\cdot(\vec{r}_i)$. The MI is a measure of the degree of correlation between $\vec{r}_i$ and $\vec{r}_i$ defined as function of $p(\vec{r}_i,\vec{r}_j)$ and $p(\vec{r}_i)\cdot(\vec{r}_i)$ accordingly to:

$$MI[\vec{r}_i,\vec{r}_j] = \int\int p(\vec{r}_i,\vec{r}_j) ln \frac{p(\vec{r}_i,\vec{r}_j)}{p(\vec{r}_i)\cdot(\vec{r}_i)} d\vec{r}_i d\vec{r}_j \tag{1.3}$$

Notably, MI is closely related to the definition of the Shannon entropy, $H[\vec{r}]$ (i.e., the expectation value of a random variable $\vec{r}$), and can be computed as:

$$MI[\vec{r}_i,\vec{r}_j] = H[\vec{r}_i] + H[\vec{r}_j] - H[\vec{r}_i,\vec{r}_j] \tag{1.4}$$

where $H[\vec{r}_i]$ and $H[\vec{r}_j]$ are the marginal Shannon entropies, and $H[\vec{r}_i,\vec{r}_j]$ is the joint entropy, providing a link between motions' correlations and information content. The g_correlation tool[92] implemented in GROMACS v4.6.4[296] was used to calculate the

marginal entropies $H[x_i]$ and $H[x_j]$ and the joint entropy $H[x_i, x_j]$ by means of the *k*-nearest neighbor distances algorithm,[297] applied to the atomic positions fluctuations from MD simulations. Since MI varies from 0 to $+\infty$, normalized generalized correlation (GC_{ij}) coefficients, ranging from 0 (independent variables) to 1 (fully correlated variables), are defined as:

$$GC_{ij}[\vec{r}_i, \vec{r}_j] = \{1 - e^{-2MI[\vec{r}_i, \vec{r}_j]/d}\}^{-1/2} \tag{1.5}$$

where d is the dimensionality of $\vec{r}_i$ and $\vec{r}_i$. Here, the GC_{ij} coefficients have been plotted as a *two–by–two* matrix; describing the correlated motions of each residue with each other. For each model system, the GC_{ij} matrices have been computed over independent MD runs, and have been averaged over the aggregate sampling arising from 4 MD replicas. The GC_{ij} matrices computed over the individual trajectories and the averaged GC_{ij} matrices are reported in Supplementary Figures A.8-A.12.

A.7 Per-domain Correlation Scores (C_s)

In order to evaluate the inter-dependent coupling between the Cas12a protein domains, we computed the per-domain generalized correlations scores (*Cs*). This measure enables to spotlight the most relevant coupled motions in large biomolecular systems, where the interpretation of correlated motions can be difficult by visual inspection of the correlation matrices. As previously shown, the per-domain *Cs* helps identifying how specific protein regions mechanistically intervene in the overall correlation network.[73,93,292] For each

protein residue i, a Cs_i parameter can be computed as:

$$Cs_i = \sum_{j \neq 1}^{N} GC_{ij} \tag{1.6}$$

which sums the GC_{ij} established by residue i with the residues j, representing a measure of the number and the intensity of the GC_{ij} coefficients displayed by each residue. To filter non-trivial correlations and eliminate the noise due to uncorrelated motions, per-residue Cs were computed considering only highly positive correlations ($GC_{ij} \geq 0.40$). To detail the overall inter-domain correlations, the Cs have been accumulated and normalized as follows. First, the Cs were calculated for each residue i belonging to a specific protein domain (e.g., Nuc), with the residues j belonging to another protein domain of interest (e.g., REC2). Then, the Cs were accumulated over all residues j of each specific Cas12a domain and normalized by the number of coupling residues. This resulted in a set of per-domain Cs, ranging from 0 (not-correlated) to 1 (correlated), measuring the strength of the overall correlation that each domain establishes with the others. As mentioned above, by using this approach we captured the most relevant correlations between spatially distant domains in large ribonucleoproteins, such as the spliceosome[93] and CRISPR-Cas9.[73] In this study, the per-domain Cs have been computed over the GC_{ij} matrices of all the FnCas12a systems averaged over 4 simulation replicas (Supplementary Figure A.13). This enabled to spotlight differences in the most relevant coupled motions of FnCas12a when bound to RNA (i.e., RNA-bound FnCas12a) and in complex with DNA, considering the effect of binding of a cleaved NTS (i.e., DNA-bound FnCas12a) and a complete NTS (i.e., DNA-bound FnCas12a').

A.8 Supplementary Figures

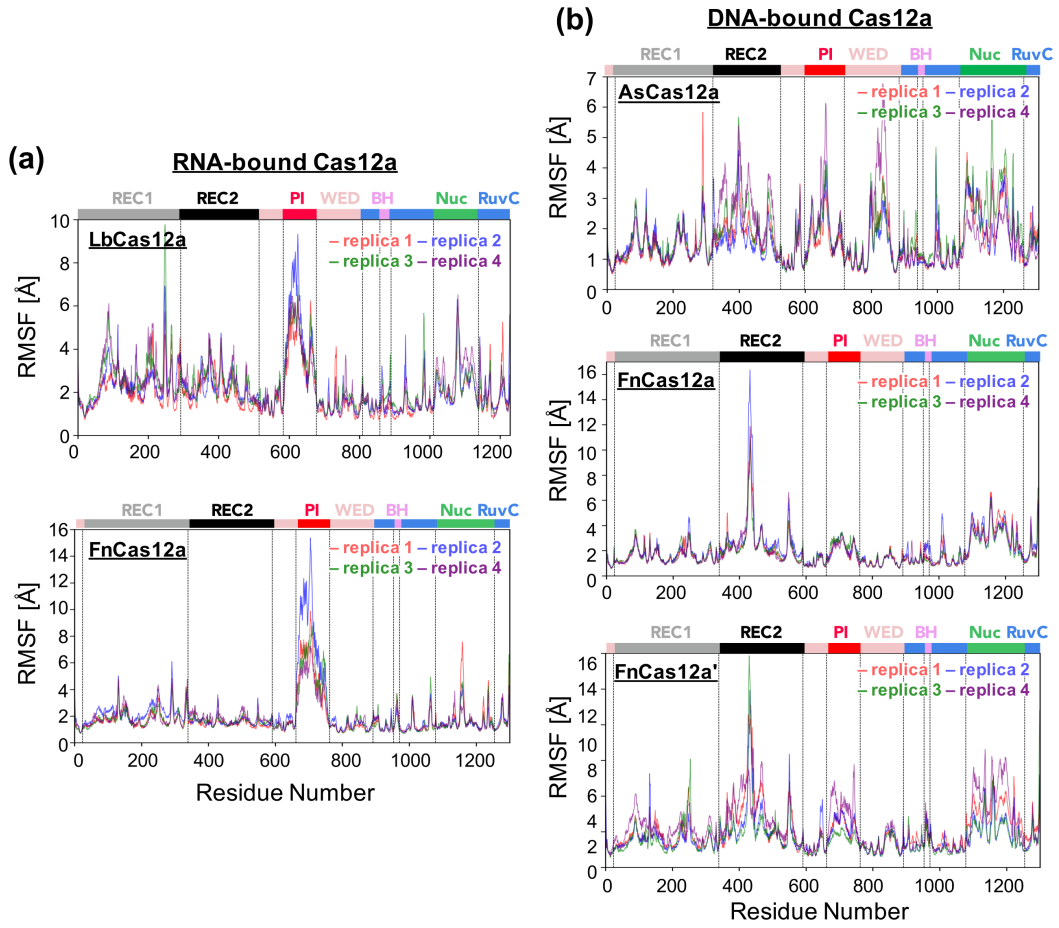


Figure A.1: **Root Mean Square Fluctuations (RMSF) profiles of the protein.** RMSF profiles are calculated for the protein Cα atoms of Cas12a bound to RNA **(a)** and DNA **(b)**. Data are reported for 5 model systems, namely the RNA-bound LbCas12a and FnCas12a; and the DNA-bound AsCas12a, FnCas12a and FnCas12a' (full details on the simulated systems are in the SI text). For each system, data are reported for 4 simulation replicas of $\sim$1 μs each. The protein sequence is shown on top of the graphs, highlighting individual protein domains in different colors.

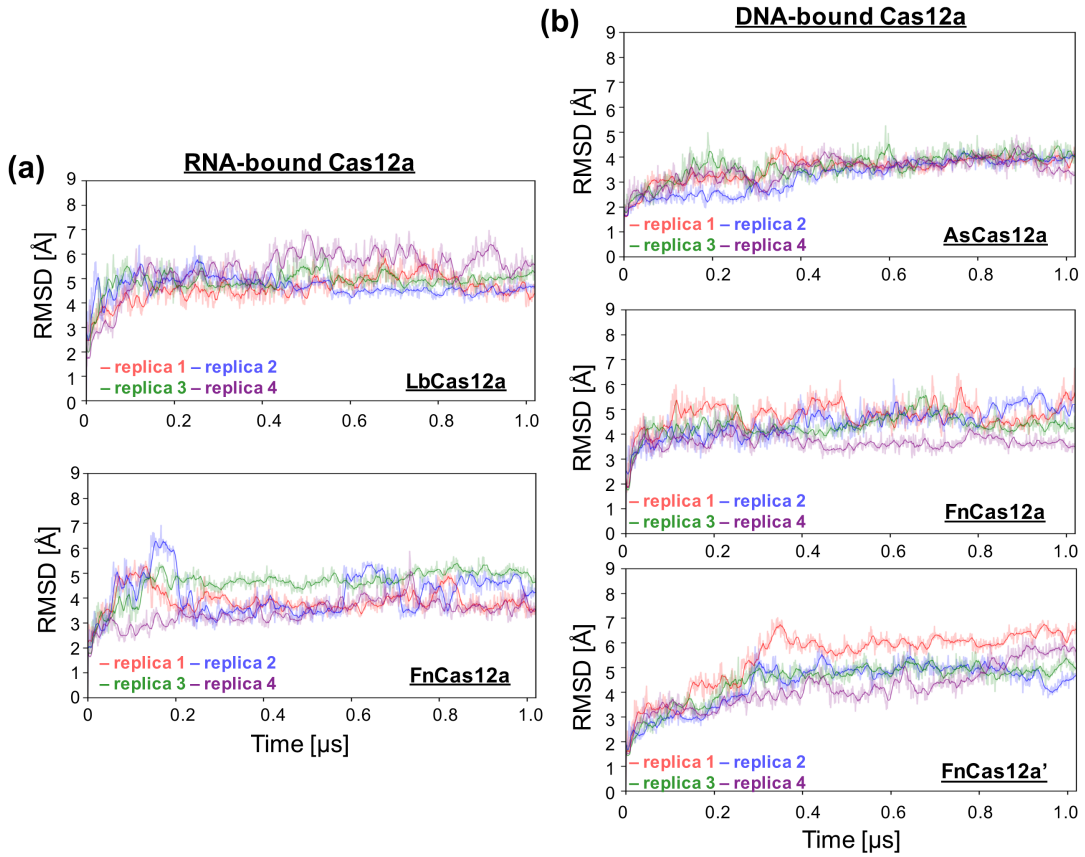


Figure A.2: **Root Mean Square Deviation (RMSD) profiles of the protein.** RMSD profiles are calculated considering the protein Cα atoms of Cas12a bound to RNA **(a)** and DNA **(b)**. Data are reported for 5 model systems, namely the RNA-bound LbCas12a and FnCas12a; and the DNA-bound AsCas12a, FnCas12a and FnCas12a' (full details on the simulated systems are in the SI text). For each system, data are reported for 4 simulation replicas of $\sim$1 μs each.

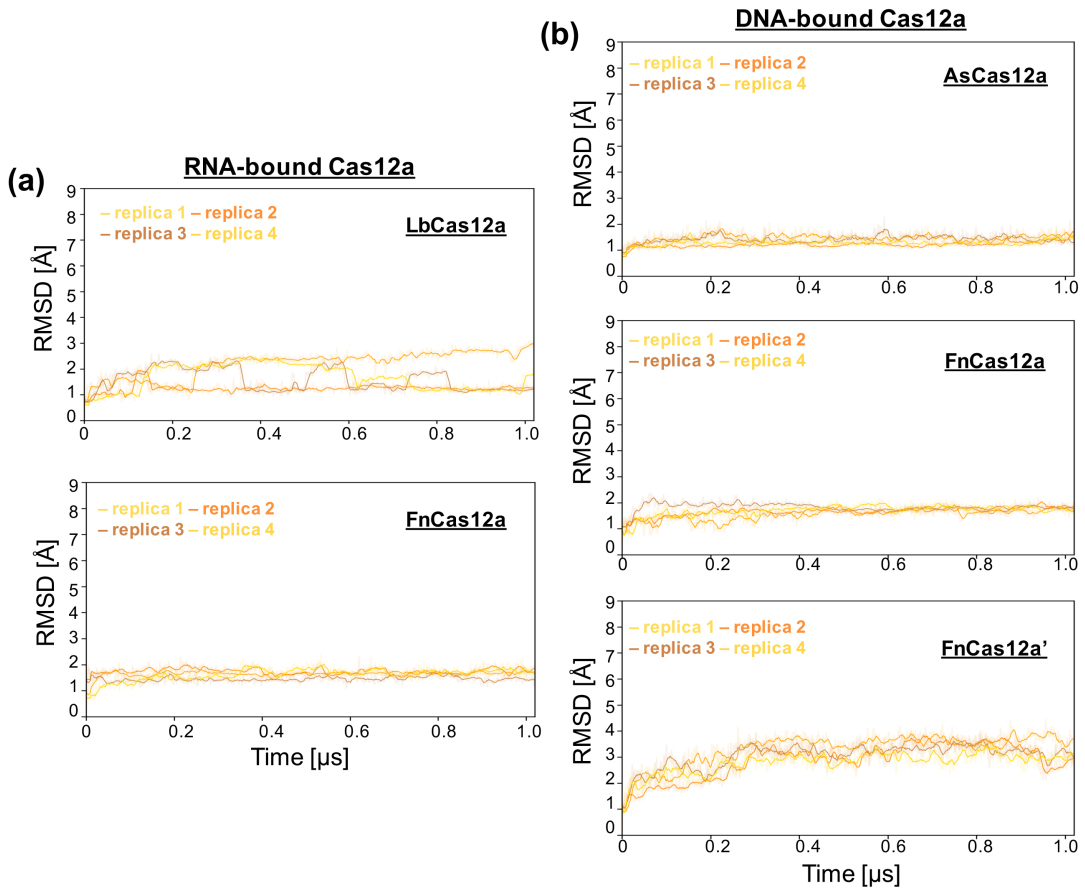


Figure A.3: **Root Mean Square Deviation (RMSD) profiles of the RNA.** RMSD profiles are calculated for the RNA backbone of Cas12a bound to RNA **(a)** and DNA **(b)**. Data are reported for 5 model systems, namely the RNA-bound LbCas12a and FnCas12a; and the DNA-bound AsCas12a, FnCas12a and FnCas12a' (full details on the simulated systems are in the SI text). For each system, data are reported for 4 simulation replicas of $\sim$1 μs each.

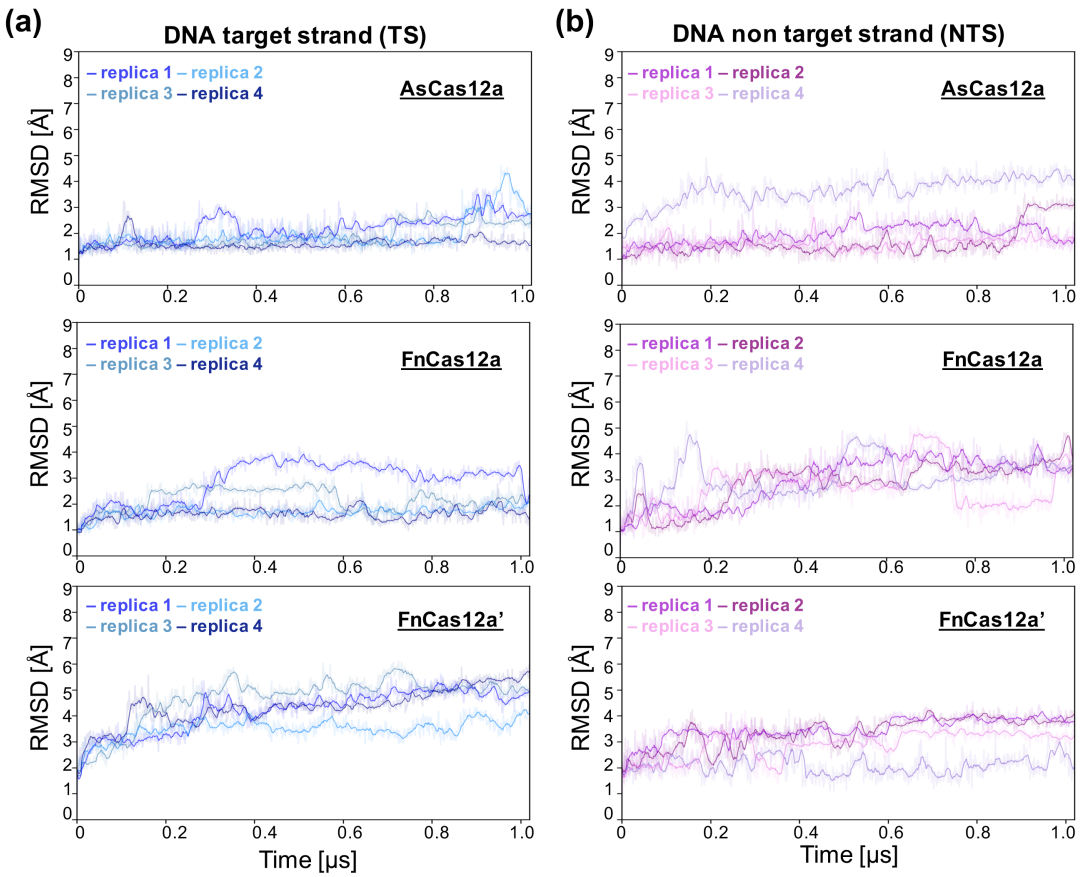


Figure A.4: **Root Mean Square Deviation (RMSD) profiles of the DNA.** RMSD profiles are calculated for the DNA target strand (TS) backbone **(a)** and for the DNA non-target strand (NTS) backbone **(b)** of Cas12a. Data are reported for 3 model systems of Cas12a bound to DNA, namely AsCas12a, FnCas12a and FnCas12a' (full details on the simulated systems are in the SI text). For each system, data are reported for 4 simulation replicas of $\sim$1 μs each.

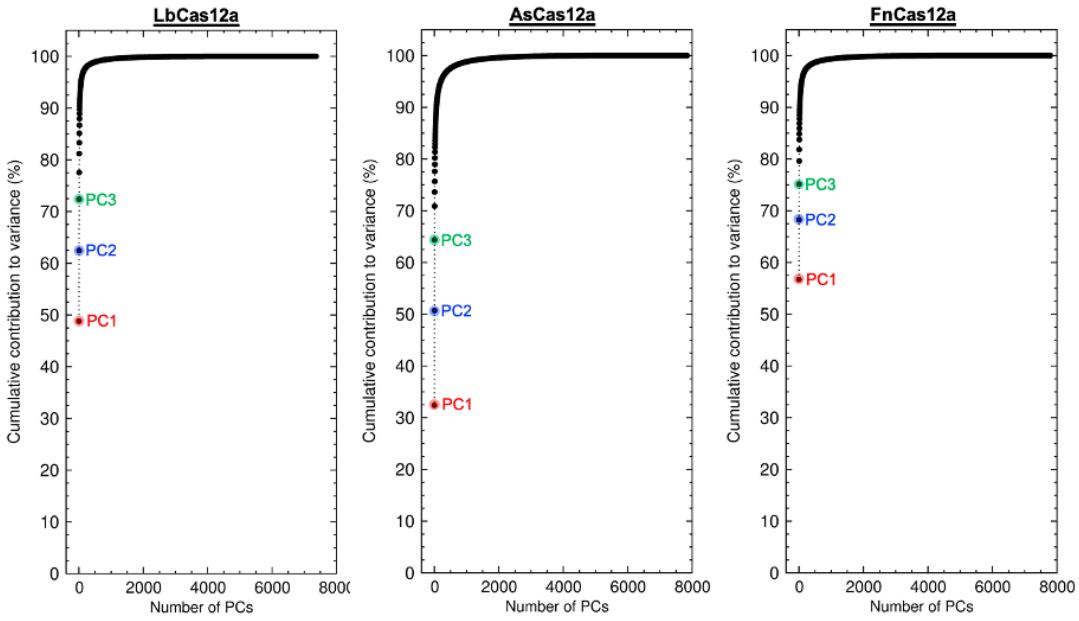


Figure A.5: **Cumulative contribution of all the principal components.** Cumulative contribution (%, y-axis) of all the principal components (PCs, *x*-axis) to the variance of the overall Cas12a motions calculated upon Principal Component Analysis (PCA) of the LbCas12a, AsCas12a and FnCas12a systems (full details on the simulated systems are in the SI text). The contribution of the first three PCs are highlighted in red, blue and green respectively.

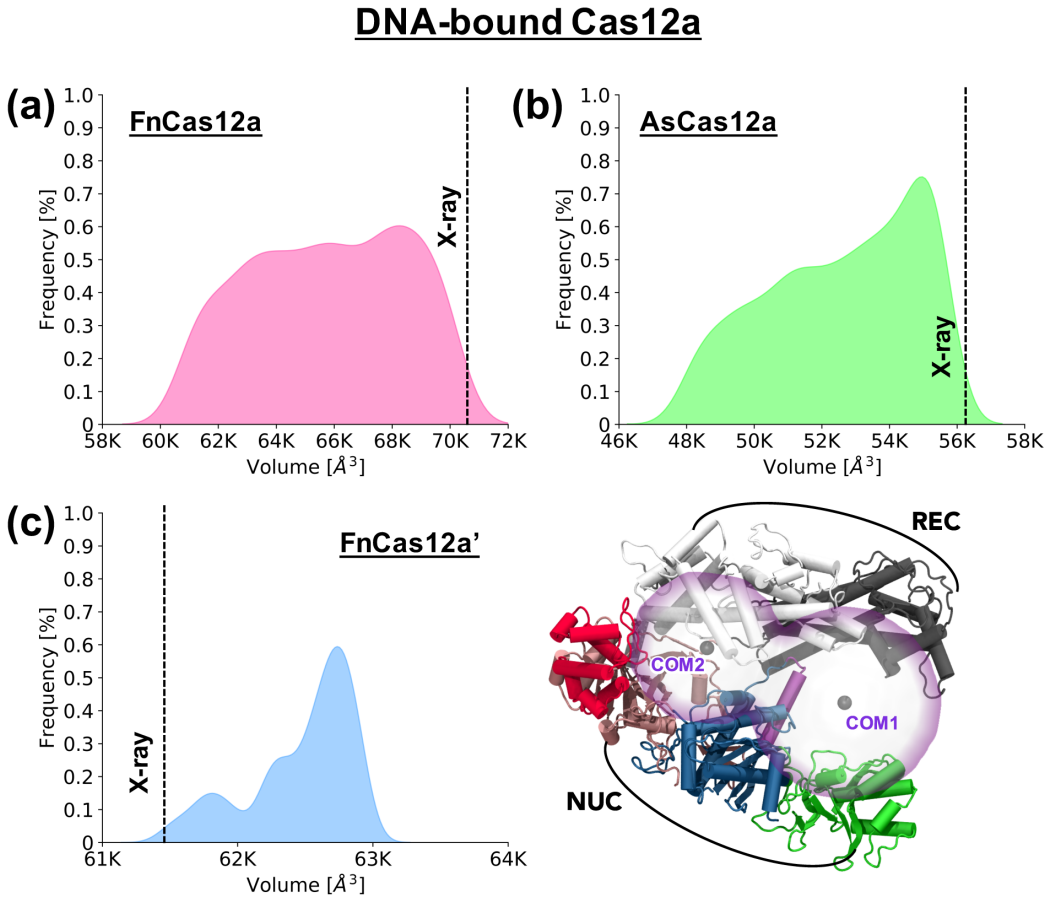


Figure A.6: **Volumetric Analysis of the groove between the REC and NUC lobes.** Probability distribution of the volume of the groove between the REC and NUC lobes, computed over the equilibrium trajectories of the DNA-bound FnCas12a **(a)**, AsCas12a **(b)** and FnCas12a' **(c)** systems. Vertical bars indicate the crystallographic values. To calculate the volume of the groove for a given trajectory frame, we monitored the space within two 20 Å radius spheres, which defined the groove between the REC and NUC lobes. The two spheres have been centered at the center of mass between the Nuc and REC2 domains (COM1) and at the center of mass between the REC, WED and RuvC (COM2) respectively, as shown in a cartoon of FnCas12a reported on the bottom right. The AsCas12a and FnCas12a' systems **(a, b)** display a contraction of the groove between the REC and NUC lobes. On the other hand, in the FnCas12a' including a complete NTS, an expansion is observed. The POcket Volume MEasurer (POVME) software (*v.* 3.0) has been used.[295] Full details on volumetric analysis are in the Supplementary Methods.

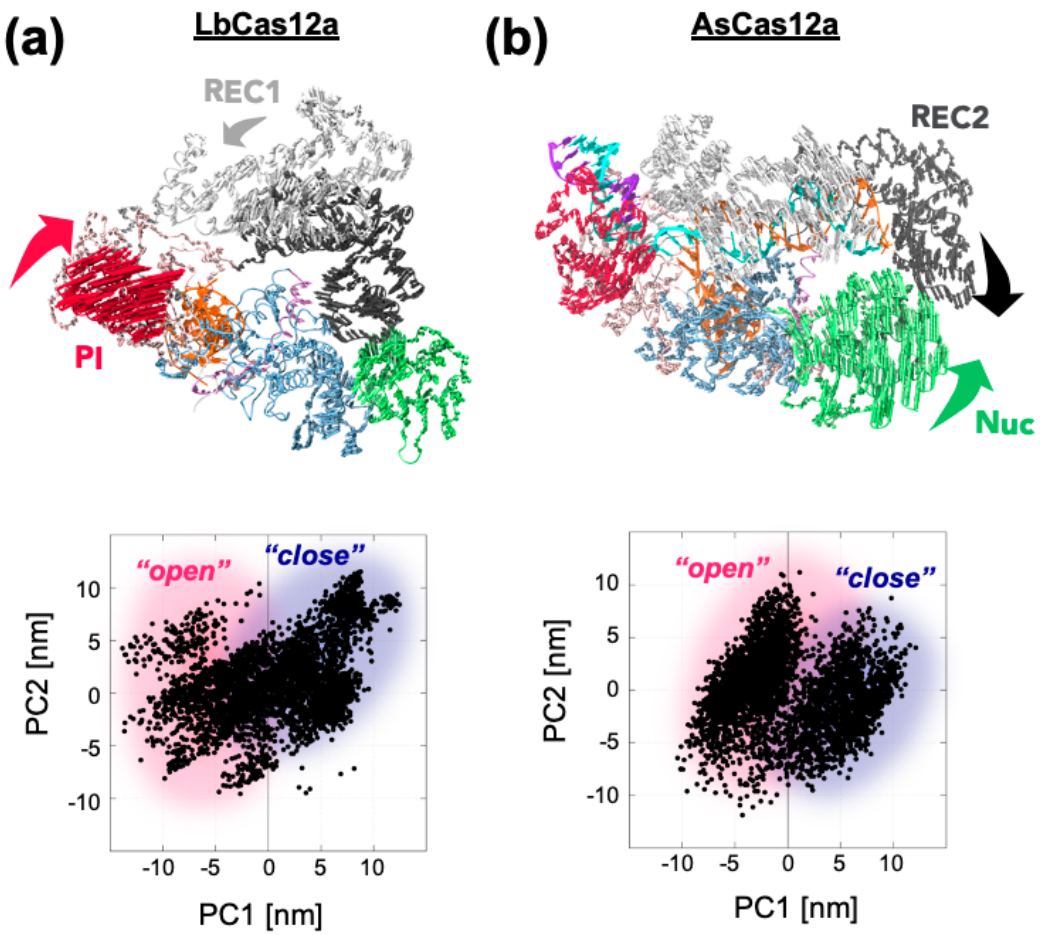


Figure A.7: **Principal Component Analysis.** **(a)** "Essential dynamics" derived from the first principal component (PC1) of the individual protein domains of the RNA-bound LbCas12a (top) and PC1 *vs.* PC2 scatter plot (bottom), characterizing the conformational space sampled. **(b)** "Essential dynamics" derived from PC1 of the individual protein domains of the DNA-bound AsCas12a (top) and PC1 *vs.* PC2 scatter plot (bottom), characterizing the conformational space sampled. Full details are reported in the Supplementary Methods.

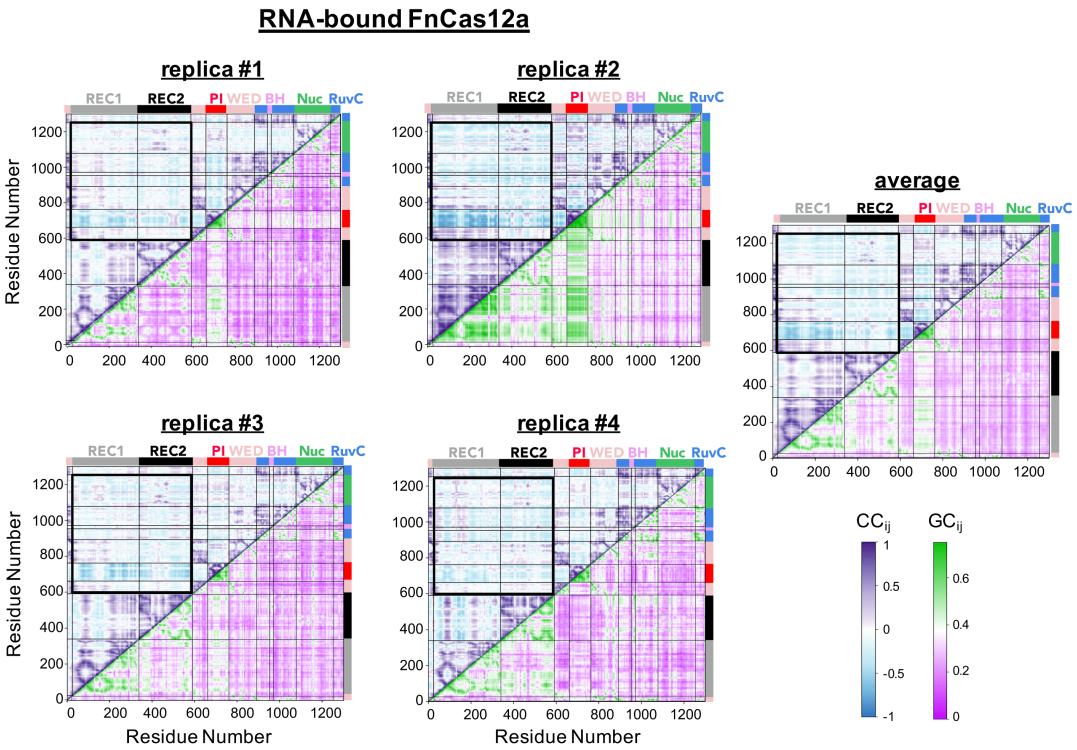


Figure A.8: **Cross-Correlation of RNA-bound FnCas12a.** Cross-Correlation (CC_{ij}, upper triangles) and Generalized Correlations (GC_{ij}, lower triangles) matrices for the RNA-bound FnCas12a. Data are reported for four independent MD simulation replicas of $\sim$1 μs each, as well as averaged over 4 simulation replicas (details are reported in the Supplementary Methods). The strength of the computed CC_{ij} is colored blue (for $CC_{ij} \geq 0$, lockstep motions) to violet (for $CC_{ij} \leq 0$, anticorrelated motions), while the GC_{ij} are color-coded green (correlated) to magenta (not correlated). The color scales are on the bottom right. The protein sequence is also shown, highlighting individual protein domains in different colors. A box is used to highlight relevant correlated motions discussed in the main text.

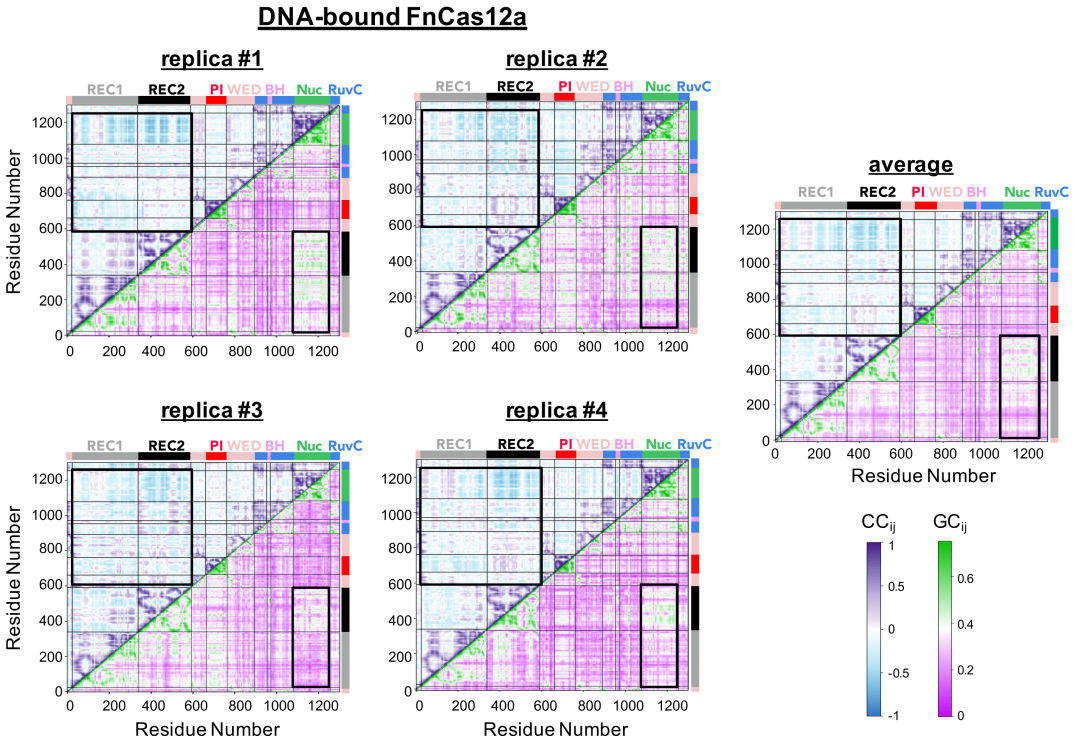


Figure A.9: **Cross-Correlation of DNA-bound FnCas12a.** Cross-Correlation (CC_{ij}, upper triangles) and Generalized Correlations (GC_{ij}, lower triangles) matrices for the DNA-bound FnCas12a. Data are reported for four independent MD simulation replicas of $\sim$1 μs each, as well as averaged over 4 simulation replicas (details are reported in the Supplementary Methods). The strength of the computed CC_{ij} is colored blue (for $CC_{ij} \geq 0$, lockstep motions) to violet (for $CC_{ij} \leq 0$, anticorrelated motions), while the GC_{ij} are color-coded green (correlated) to magenta (not correlated). The color scales are on the bottom right. The protein sequence is also shown, highlighting individual protein domains in different colors. A box is used to highlight relevant correlated motions discussed in the main text.

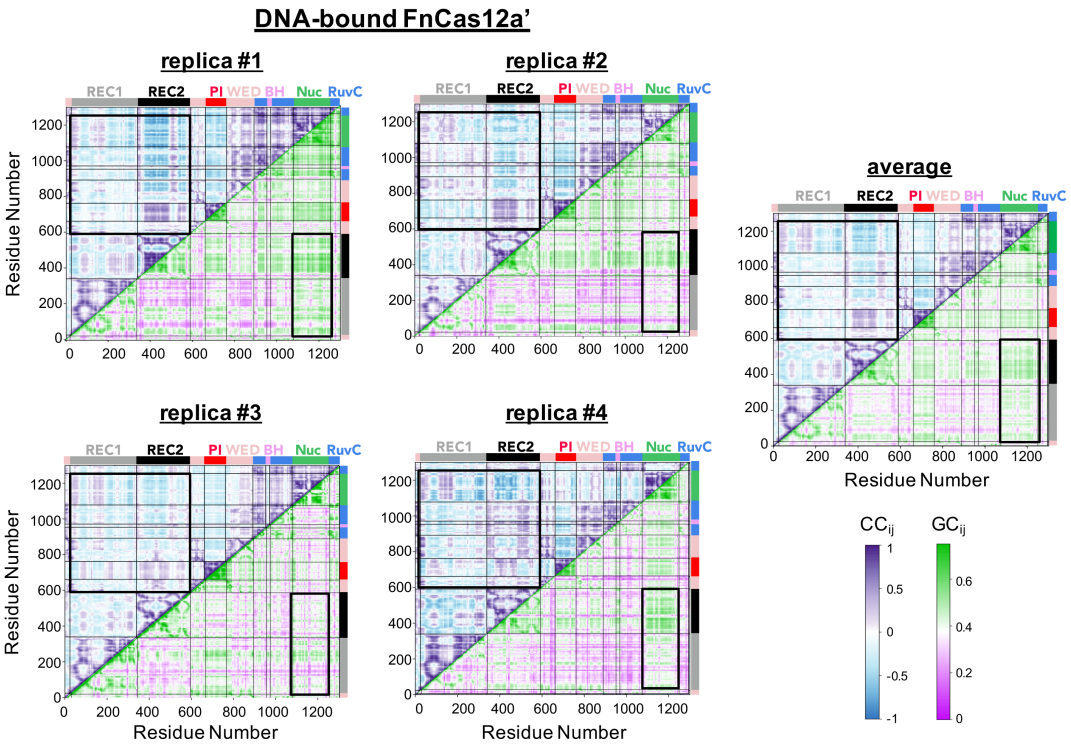


Figure A.10: **Cross-Correlation of DNA-bound FnCas12a'.** Cross-Correlation (CC_{ij}, upper triangles) and Generalized Correlations (GC_{ij}, lower triangles) matrices for the DNA-bound FnCas12a' (including a complete NTS). Data are reported for four independent MD simulation replicas of $\sim$1 μs each, as well as averaged over 4 simulation replicas (details are reported in the Supplementary Methods). The strength of the computed CC_{ij} is colored blue (for $CC_{ij} \geq 0$, lockstep motions) to violet (for $CC_{ij} \leq 0$, anticorrelated motions), while the GC_{ij} are color-coded green (correlated) to magenta (not correlated). The color scales are on the bottom right. The protein sequence is also shown, highlighting individual protein domains in different colors. A box is used to highlight relevant correlated motions discussed in the main text.

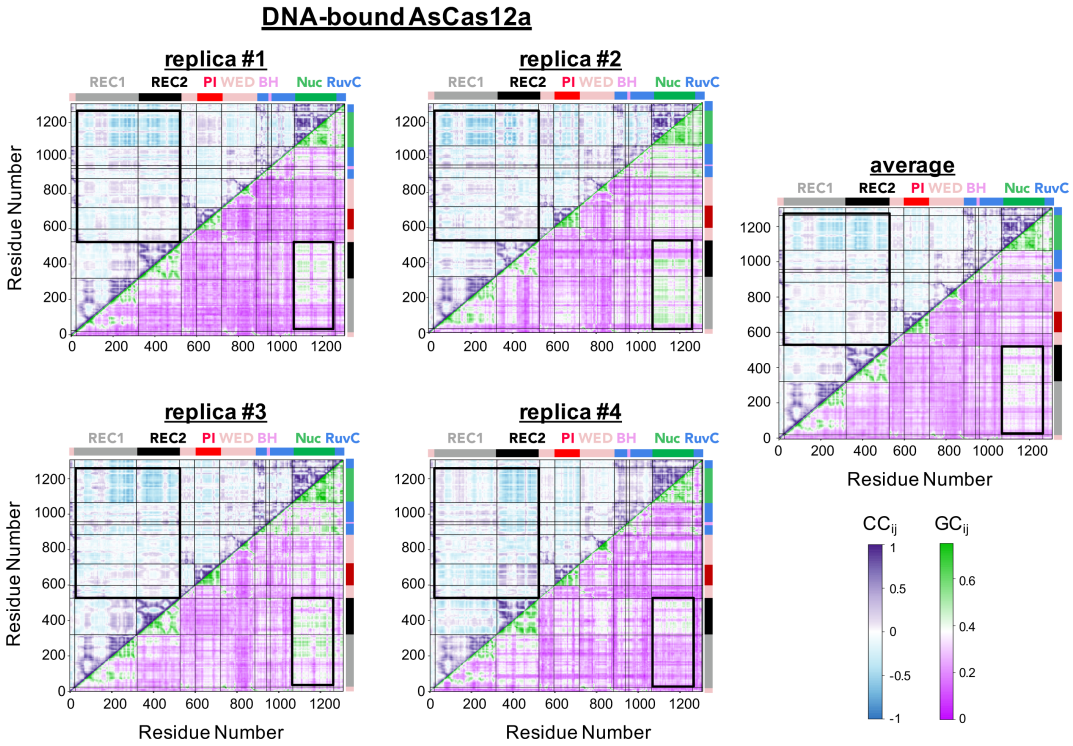


Figure A.11: **Cross-Correlation of DNA-bound AsCas12a.** Cross-Correlation (CC_{ij}, upper triangles) and Generalized Correlations (GC_{ij}, lower triangles) matrices for the DNA-bound AsCas12a. Data are reported for four independent MD simulation replicas of $\sim$1 μs each, as well as averaged over 4 simulation replicas (details are reported in the Supplementary Methods). The strength of the computed CC_{ij} is colored blue (for $CC_{ij} \geq 0$, lockstep motions) to violet (for $CC_{ij} \leq 0$, anticorrelated motions), while the GC_{ij} are color-coded green (correlated) to magenta (not correlated). The color scales are on the bottom right. The protein sequence is also shown, highlighting individual protein domains in different colors. A box is used to highlight relevant correlated motions discussed in the main text.

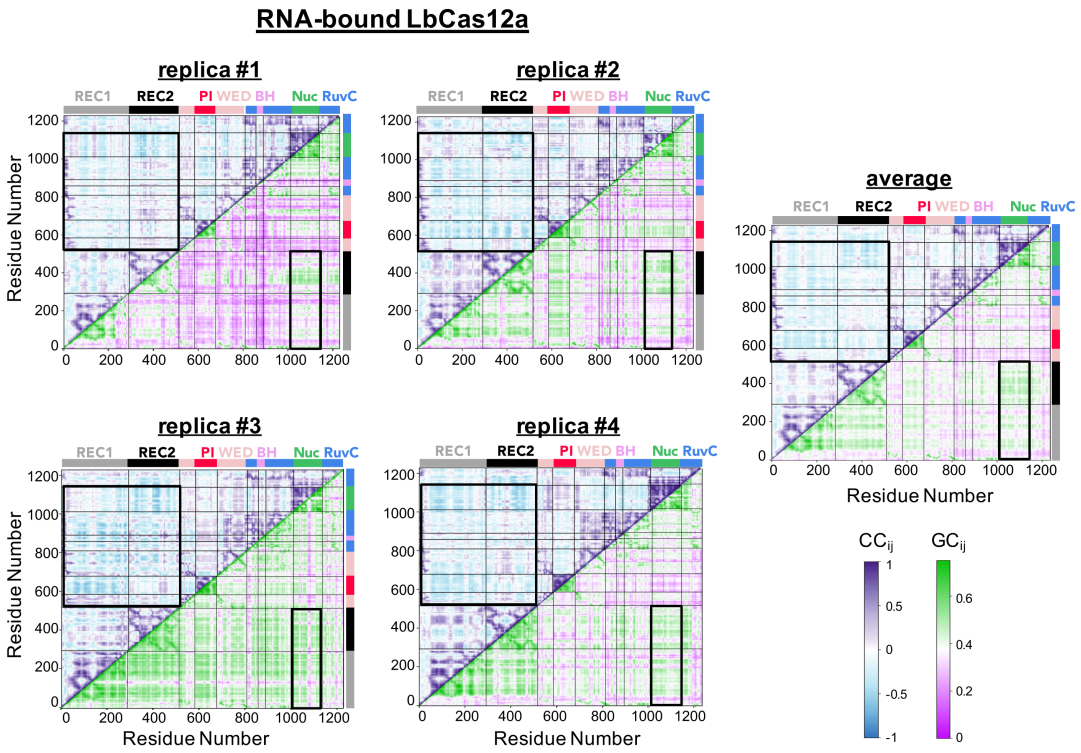


Figure A.12: **Cross-Correlation of RNA-bound LbCas12a.** Cross-Correlation (CC_{ij}, upper triangles) and Generalized Correlations (GC_{ij}, lower triangles) matrices for the RNA-bound LbCas12a. Data are reported for four independent MD simulation replicas of $\sim$1 μs each, as well as averaged over 4 simulation replicas (details are reported in the Supplementary Methods). The strength of the computed CC_{ij} is colored blue (for $CC_{ij} \geq 0$, lockstep motions) to violet (for $CC_{ij} \leq 0$, anticorrelated motions), while the GC_{ij} are color-coded green (correlated) to magenta (not correlated). The color scales are on the bottom right. The protein sequence is also shown, highlighting individual protein domains in different colors. A box is used to highlight relevant correlated motions discussed in the main text.

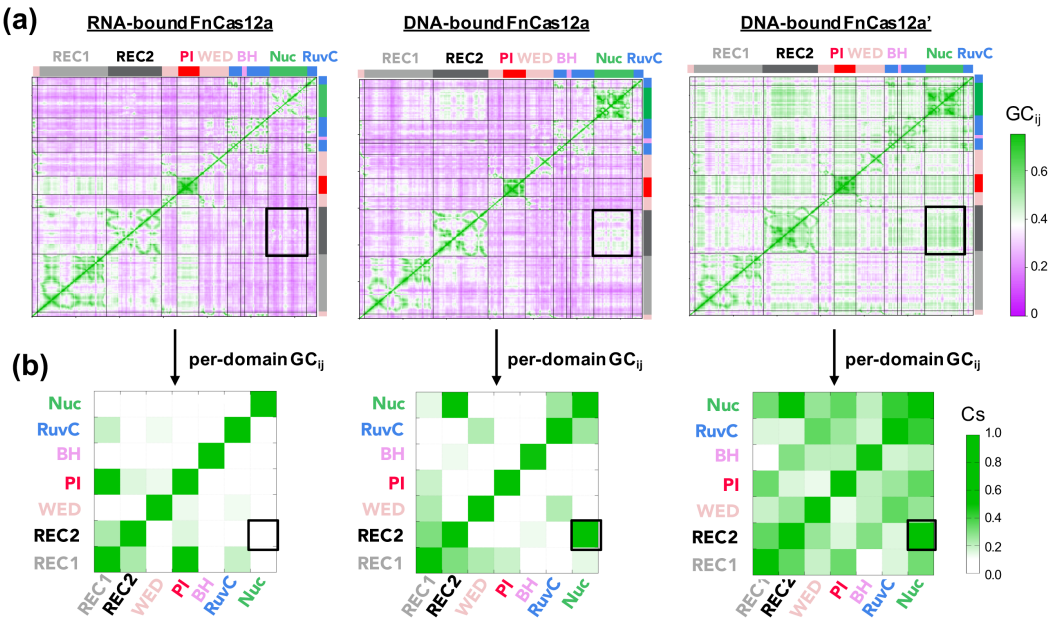


Figure A.13: **Generalized Correlation Analysis.** **(a)** Generalized Correlations (GC_{ij}) matrices for FnCas12a bound to RNA (i.e., RNA-bound FnCas12a, left panel) and in complex with DNA, binding a cleaved NTS (i.e., DNA-bound FnCas12a, central panel) and a complete NTS (i.e., DNA-bound FnCas12a, right panel). For each system, the GC_{ij} matrices are averaged over 4 simulation replicas of 1 μs each. The strength of the computed GC_{ij} is colored from green (correlated) to magenta (not correlated, color scale on the right). The protein sequence is also shown. **(b)** Per-domain GC_{ij} Score (C_s) matrices, which accumulate (and normalize) the per-domain GC_{ij}. The per-domain C_s matrices are a "coarse" representation of the GC_{ij} matrices (reported in panel **(a)**), which can spotlight the most relevant correlations between spatially distant domains, ranging from 0 (not-correlated, white) to 1 (correlated, green). Upon DNA binding, Nuc and REC2 display the highest C_s (a box is used to indicate the REC2/Nuc region on the correlation matrices, highlighting the absence of correlations prior DNA binding). Full details are reported in the Supplementary Methods.

Appendix B

Supporting Information for Chapter III

B.1 Structural Models

Molecular simulations were based on two structures of the *F. novicida* Cas12a obtained upon NTS cleavage: (1) the cryo-EM structure EMD-0065 (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation))[1] solved at 3.27 Å resolution and (2) the X-ray structure PDB: 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop),[2] solved at 2.50 Å resolution. The 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) cryo-EM structure contains an α-helical "lid" (residues L1008 – K1021, Figure 3.1a), majority of which is structurally disordered in the 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) X-ray structure. Hence, the missing N1009 – F1017 residues in the 5NFV

[https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) structure, along with other missing loops, were reconstructed through homology modelling using the SWISS-MODEL software[125,289] (Figure 3.1b), where we used the X-ray 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) structure as the template.[125]

We built the complete Cas12a system (Figure 3.1c) starting from the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation)[1] cryo-EM structure of Cas12a (Figure 3.1a), in which the lid displays an α-helical form. Since 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) lacks the 5'-tail of the TS (bases from -1 to -7) downstream of the crRNA:TS duplex, containing the scissile phosphate,[58] we built it from the 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) X-ray structure of Cas12a (Figure 3.1b) that contains a longer TS. In detail, we performed an all-Cα based RMSD fitting of the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) structure on 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) (RMSD = ∼2 Å, Supplementary Figure B.2a). Next, we performed a separate backbone based RMSD fitting (RMSD = ∼2.5 Å) of the nucleic acids between the two structures and patched the TS:NTS hybrid downstream of the crRNA:TS duplex from the 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) structure onto 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation). We maintained the nucleic acid sequence of the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) structure. This complex was further refined using the multi-template mode of the program MODELLER.[298] The RuvC domain of Cas12a

is highly conserved across nucleases,[299] including the Cas9 enzyme[115] that was first harnessed for gene editing.[8] It holds a two-metal aided architecture, characterized by the highly conserved DEDD (or DDE) motif,[2] in which Mg^{2+} ions are coordinated by carboxylate groups. The X-ray structure of Cas12i2 (PDB: 6LTU [https://doi.org/10.2210/pdb6LTU/pdb] (Cas12i2 ternary complex)[169]), phylogenetically related to Cas12a, captured the catalytic Mg^{2+} ions coordinating the DED motif in RuvC (Supplementary Figure B.2b). The two catalytic Mg^{2+} ions were located in the RuvC site of Cas12a by superimposing the DDE motif of the two structures (6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) and 6LTU [https://doi.org/10.2210/pdb6LTU/pdb] (Cas12i2 ternary complex)). Finally, the complete Cas12a system was subjected to force field-based vacuum minimization to obtain an energetically stable complex, further used for long timescale MD simulations. Importantly, distance analyses of the DDE motif coordinating with the Mg2+ ions along $\sim$10 μs-long MD simulation (Supplementary Figure B.2c) underscores the stability of the two-metal ion architecture in the system.

The catalytically competent state of Cas12a for TS cleavage was built taking cues from related crystal structures of Cas12a orthologues, holding the DNA TS within the RuvC active site. The X-ray structure of the CRISPR-Cas12b (PDB: 5U30 [https://doi.org/10.2210/pdb5U30/pdb] (AacC2c1-sgRNA-extended target DNA ternary complex),[172] at 2.9 Å resolution) accommodates the TS within the RuvC catalytic cleft (Supplementary Figure B.8a). The TS bends down from the crRNA:TS duplex toward the active site, forming a U-shaped strand with its 5'-tail interacting with the Nuc domain. Cas12a and Cas12b share a highly conserved RuvC catalytic domain, which is

superposable in the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) and 5U30 [https://doi.org/10.2210/pdb5U30/pdb] (AacC2c1-sgRNA-extended target DNA ternary complex) structures (Supplementary Figure B.8b). A comparison of the two structures shows that the nucleic acids are also superposable and differ for a longer U-shaped TS that locates with the RuvC active site in Cas12b. Building on this similarity, the 5U30 [https://doi.org/10.2210/pdb5U30/pdb] (AacC2c1-sgRNA-extended target DNA ternary complex) structure was used as a template for biased MD simulations, starting from the complete Cas12a system that includes Mg^{2+} ions in the RuvC site. Restraint forces were applied to reduce the RMSD between initial and final conformations of the DNA nucleobases -1 to -5, which include the scissile phosphate, similar to a prior modeled structure of Cas12a.[2] The following potential was applied:

$$\begin{aligned} V = \frac{k}{2N}(RMS(t) - RMS^*(t)), \quad RMS(t) > RMS^*(t) \\ V = 0, \quad RMS(t) \leq RMS^*(t) \end{aligned} \tag{2.1}$$

where k is the collective force constant, N is the number of targeted atoms, $RMS(t)$ is the actual RMSD between the current (at time t, initially) and target configurations, and $RMS^*(t)$ is reduced with a constant velocity from the initial RMSD value between the initial and target configurations toward the desired final RMSD (i.e., zero). The difference in RMSD was gradually reduced in 30 steps of 20 ns each with a force constant k of 30 kcal mol^{-1} $Å^{-2}$, collecting an overall ensemble of ∼600 ns. This protocol leads to a relaxed trajectory, allowing the protein to adapt and accommodate the transition. The resultant structure contains the scissile phosphate, between the nucleobases T_{-2} and G_{-3}, which coordinates the catalytic Mg^{2+} ions (Supplementary

Figure B.8c,d). This complex was subjected to ∼10 μs of MD simulation for further refinement and to ensure long timescale stability (Supplementary Figure B.9). Along the simulations, the active site is stable (Supplementary Figure B.9c) and the DDE motif steadily coordinates Mg^{2+} (Supplementary Figure B.9d).

All the systems were embedded in explicit waters, while counterions were added to neutralize the total charge at physiological conditions, leading to periodic simulation cells of ∼138*149*167 $\mathring{A}^3$ and a total of ∼307,000 atoms for each system.

B.2 Molecular Dynamics (MD) Simulations

Conventional MD simulations were performed to equilibrate the systems prior to long timescale MD on Anton-2.[114] A simulation protocol tailored for RNA/DNA nucleases was adopted, embracing the use of the Amber ff19SB[126] force field, which includes the ff99bsc1 corrections for DNA[127] and the ff99bsc0+χOL3 corrections for RNA.[101,102] The TIP3P model was employed for explicit water molecules.[104] The Li & Merz 12-6 model of non-bonded interactions was used for Mg^{2+} ions.[128] We have extensively employed these force field models in computational studies of CRISPR-Cas systems,[129] showing also that they perform well for long timescale simulations on Anton-2.[75] The Li & Merz model also reported a good description of Mg^{2+} bound sites, in agreement with quantum/classical simulations.[130] An integration time step of 2 fs was employed. All bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm. Temperature control (300 K) was performed via Langevin dynamics,[107] with a collision frequency γ = 1/ps. Pressure control was accomplished by coupling the system to a

Berendsen barostat[108] at a reference pressure of 1 atm and with a relaxation time of 2 ps. The systems were subjected to energy minimization to relax water molecules and counter ions, keeping the protein, the RNA, DNA and Mg^{2+} ions fixed with harmonic position restraints of 300 kcal/mol $\cdot$ $\mathring{A}^2$. Then, the systems were heated up from 0 to 100 K in a canonical ensemble (NVT), by running two simulations of 5 ps each, imposing position restraints of 100 kcal/mol $\cdot$ $\mathring{A}^2$ on the above-mentioned elements of the system. The temperature was further increased up to 200 K in $\sim$100 ps of MD in the isothermal-isobaric ensemble (NPT), reducing the restraint to 25 kcal/mol $\cdot$ $\mathring{A}^2$. Subsequently, all restraints were released, and the temperature of the systems was raised up to 300 K in a single NPT simulation of 500 ps. After $\sim$ 1.1 ns of equilibration, $\sim$10 ns of NPT runs were carried out allowing the density of the systems to stabilize around 1.01 g cm^{-3}. Finally, production runs were carried out in the NVT ensemble, collecting $\sim$120 ns. These simulations have been performed using the GPU-empowered version of AMBER 20.[131] The well-equilibrated systems were used as starting points for simulations on Anton-2.[114]

Long timescale MD simulations of the CRISPR-Cas12a systems were performed using Anton-2,[114] a special-purpose supercomputer for micro-to-millisecond length MD simulations, starting from well-equilibrated configurations, obtained after $\sim$120 ns of conventional MD (*vide supra*). Simulations on Anton-2 were performed using the same force field parameters used for conventional MD simulations. A reversible multiple time step algorithm[132] was employed to integrate the equations of motion with a time step of 2 fs for short-range non-bonded and bonded forces and 6 fs for the long-range

non-bonded forces. Simulations were performed in the NPT ensemble using the multigrator integrator as implemented in Anton-2.[133] Pressure control was accomplished via the Martyna, Tobias, Klein (MTK) barostat,[134] set to maintain 1 bar of pressure, with a tau (piston time constant) parameter of 0.0416667 ps and reference temperature of 310.15 K. The barostat period was set to the default value of 480 ps per timestep. Temperature control was accomplished via the Nosé-Hoover thermostat[135,136] with a tau (time constant) parameter of 0.0416667 ps. The k-Gaussian split Ewald method[137] was used for long-range electrostatic interactions. Hydrogen atoms were added assuming standard bond lengths and were constrained to their equilibrium position with the SHAKE algorithm.[138] By using this approach and the Anton-2 supercomputer,[114] each model system was simulated for $\sim$10 μs and in replicates. Specifically, we obtained multiple trajectories of the 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) cryo-EM structure (simulated for $\sim$10 μs) and its more complete model including a longer DNA ($\sim$10 μs in two replicates), as well as of the 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) X-ray structure ($\sim$10 μs in two replicates). The structural model of the catalytically competent Cas12a including the DNA TS within the RuvC site was also simulated for $\sim$10 μs. This resulted in a total of $\sim$60 μs of all-atom classical MD simulations.

B.3 Principal Component Analysis

PCA is a statistical method used to reduce a large number of degrees of freedom to an essential subspace set, which reports on the large-amplitude, collective motions in

biological molecules undergoing MD simulations. In PCA, the covariance matrix of the protein Cα atoms is calculated and diagonalized to obtain a new set of coordinates (eigenvectors) to describe the system motions. Each eigenvector – also called Principal Component (PC) – is associated with an eigenvalue corresponding to the mean square fluctuation contained in the system's trajectory projected along that eigenvector. By sorting the eigenvectors according to their eigenvalues, the first PC1 corresponds to the system's largest amplitude motion, and the dynamics of the system along PC1 are usually referred "essential dynamics".[87,291]

Here, PCA has been performed considering the simulations of the complete Cas12a after NTS cleavage (Figure 3.1c, Table B.1). First, the collective ensembles were combined and subjected to RMS-fit to the same reference configuration, removing rotational and translational motions, to ensure a consistent eigenbasis on all compared systems. Then, the principal motions of the protein have been captured starting from the mass-weighted covariance matrix of the Cα atoms. Each element in this matrix is the covariance between atoms i and j, defining the i, j position of the covariance matrix (C_{ij}):

$$C_{ij} = \langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle \tag{2.2}$$

where $\vec{r}_i$ and $\vec{r}_i$ are the position vectors of atoms i and j, and the brackets denote an average over the sampled time period. A positive sign of this product indicates that the two atoms move in a correlated manner, otherwise, a negative value points to anti-correlated atoms. If the product is zero, then it evinces that the atoms' displacements are independent of each other. The covariance matrix was then diagonalized, leading to a complete set of orthogonal collective eigenvectors, each associated with a corresponding

eigenvalue. The eigenvalues denote how much each eigenvector is representative of the system dynamics, thus giving a measure of the contribution of each eigenvector to the total variance. Indeed, the eigenvectors with the largest eigenvalues correspond to the most relevant motions. By projecting the displacement vectors of each atom along the trajectory onto the eigenvectors (i.e., by taking the dot product between the two vectors at each frame), the PCs were obtained. The first PC1, commonly referred to as "essential dynamics",[87,291] is plotted on the 3D structure of the complete CRISPR-Cas12a system indicating the largest amplitude motions of the protein (Supplementary Figure B.6e). PCA has been performed using cpptraj of AMBER 20,[131] while the Normal Mode Wizard[293] plugin of the Visual Molecular Dynamics[294] program has been used for the graphical rendering.

B.4 Umbrella Sampling Simulations

The umbrella sampling (US) method[117] was used to compute the free energy profile associated with the traversal of the DNA TS toward the RuvC active site. In this method, several simulations (US windows) are run in parallel with additional harmonic bias potential applied to selected Reaction Coordinates (RCs):

$$V(RC) = \frac{k}{2}(RC(t) - RC^*)^2 \tag{2.3}$$

where $V(RC)$ is the value of the bias potential, k is a bias force constant, $RC(t)$ is the value of RC at given time t and RC^* is the reference value of RC. By using different RC^* values in each US window, one can sample the biased probability distribution

$p_b(RC)$ along the whole RC range of interest. Thanks to a simple reweighting scheme providing the unbiased probabilities, the free energy of a process along the RCs can be computed (*vide infra*). We performed two-dimensional Umbrella Sampling (2-D US) simulations at 300 K using two reaction coordinates (RCs). *RC*1 was the difference in root-mean-square deviation (RMSD) of all the Cα atoms between the initial (Figure 3.1c) and the final (Supplementary Figure B.8d) states. *RC*2 was the distance between the centre of mass of the RuvC catalytic core (D917, E1006, D1255 coordinating with the two catalytic Mg^{2+} ions) and the TS region including the scissile phosphate (i.e., the DNA nucleobases at positions -2 to -4, as indicated by biochemical and single-molecule experiments).[53,58] Restraint forces were applied on the two RCs as follows. The difference in RMSD between initial and final conformations was reduced, according to the following potential:

$$\begin{aligned} V = \frac{k}{2N}(RMS(t) - RMS^*(t)), \quad RMS(t) > RMS^*(t) \\ V = 0, \quad RMS(t) \leq RMS^*(t) \end{aligned} \tag{2.4}$$

where k is the collective force constant, N is the number of targeted atoms, $RMS(t)$ is the actual RMSD between the current (at time t, initially) and target configurations, and $RMS^*(t)$ is reduced with a constant velocity from the initial RMSD value ($\sim$6 Å) between the initial and target configurations toward the desired final RMSD (i.e., zero). *RC*1 was divided into bin sizes of 0.2 Å RMSD difference, and each window was run for $\sim$10 ns with a collective force constant k of 50 kcal mol^{-1} $Å^{-2}$. Additionally, restraint forces were also applied to reduce the *RC*2 from $\sim$30 Å to $\sim$6.4 Å, according to the same equation. *RC*2 was discretized into bin sizes of 0.2 Å and individual windows were

run for ∼15 ns with a collective force constant k of 50 kcal mol^{-1} Å^{-2}. Overall, we simulated 28 ($RC1$) * 119 ($RC2$) = 3,332 windows, resulting in a collective ensemble of ∼50 μs simulations. Two independent sets of 2-D US simulations were performed: (i) for the wild-type system and (ii) upon mutating relevant residues of the α-helical lid (N1009, F1012, K1013, R1014, R1016, K1018) into alanine. Then, the free energy profiles were computed using the Weighted Histogram Analysis (WHAM) method[139] upon removing the initial 1/3rd of the US trajectories from each window, corresponding to the relaxation of the system. The unbiased probability distribution ($p_u(RC)$) was recovered from the biased probability distribution ($p_b(RC)$) through the formula:

$$p_u(RC) = p_b(RC) * exp(\frac{V(RC) - F}{k_B T}) \tag{2.5}$$

where $V(RC)$ is the value of the applied potential at given RC in a given US window, k_B is the Boltzmann constant, T is the reference temperature, F corresponds to an expected value of $exp^{V(RC)}$ for each US window and is optimized with the WHAM algorithm.[139] Analysis of the conformational ensembles was performed on the reweighted trajectories using the exponential factor $exp(\frac{V(RC)-F}{k_B T})$.

Analysis of the free energy landscape was performed by computing the minimum free energy pathway using an approach similar to the Dijkstra algorithm.[140] In detail, the map is sampled from the origin point, propagating to the lowest energy point available on each iteration until it reaches the target point. This enabled us to consider both the RCs along a minimum free energy path. The convergence of the free energy profiles was evaluated by plotting the minimum free energy pathways, considering the remaining 10

ns runs from each US window, for ∼5 ns through ∼10 ns (Supplementary Figure B.12). The convergence is reached ∼8 ns onward for both systems and hence we used bin sizes of ∼10 ns to compute our 2-D free energy surfaces (Supplementary Figure B.11) and the minimum free energy paths (Figure 3.4 and Supplementary Figure B.11). The error estimation on the minimum free energy pathways was performed using the Monte Carlo bootstrap error analysis.[141]

B.5 Calculation of Bending Angles

Conformational changes of the protein and nucleic acids were monitored through four critical bending angles describing: (i) crRNA:TS duplex arching, (ii) Rec2 bending and (iii) Nuc bending, (iv) TS bending. In cases (i), (ii) and (iii), we computed the angle between two vectors passing through two regions of interest (defined as mask 1 and mask 2, Table B.2). We generated a plane based on the first frame of MD production runs by fitting the Cα atoms of the residues under selection (or P atoms for the nucleobases) with the smallest mean square error. A vector ($\vec{V}$) was generated through the centre of mass of two critical points of the selection relevant to the conformational change of interest (viz., mask1 to mask2, Table B.2) on the plane. Another vector ($\vec{V_{\perp}}$), perpendicular to $\vec{V}$, out of the plane was generated from the frame of reference. The bending angles were studied by calculating the angle of the vector ($\vec{V}$) along the trajectory with respect to $\vec{V_{\perp}}$. For the calculation of (iv) TS bending, we calculated the angle between two vectors ($\vec{V_1}$ and $\vec{V_2}$) passing through the centre of mass of two critical points of the selection relevant to the conformational change of interest (viz., mask1 to mask2, Table B.2).

B.6 Alchemical Free Energy Simulations

We computed the relative free energy changes (i.e., $\Delta\Delta G$) of the individual residues with respect to alanine in binding the DNA target strand (TS), by performing alchemical free energy simulations.[142] Here, the DNA target strand (TS, ligand) bound to the Cas12a complex (receptor) through the α-helical lid was decoupled to an unbound state following the alchemical free energy formalism using softcore (SC) potentials,[121] by gradually diminishing the electrostatic and van der Waals interactions of the TS (A_{14}, C_2, G_1 and G_{-1}) bases that interact with the lid. We used a lambda (λ) dynamic variable to define the thermodynamic states of the system along this decoupling alchemical pathway. Thermodynamic integration (TI) simulations were performed at 21 λ values from 0 to 1, at an interval of 0.05, for the wild-type (WT) CRISPR-Cas12a and for each individual alanine mutant. The systems were minimized and equilibrated using the simulation protocol described above. The final binding free energies were computed using the Multistate Bennett Acceptance Ratio (MBAR) method[118,119] to integrate the free energies over the different lambda values.[143,144] The relative binding free energy ($\Delta\Delta G$) was calculated as follows:

$$\Delta\Delta G_{residue} = \Delta G_{WT} - \Delta G_{mutated} \tag{2.6}$$

In detail, the pairwise interactions of the DNA TS region (A_{14}, C_2, G_1 and G_{-1}) responsible for binding the lid (L1008 to K1021) was "softened" by modifying the Lenard-Jones (LJ) softcore potentials for van der Waal interactions and Coulombic

(Coul) softcore potentials for electrostatic interactions as follows:

$$U_{SC}^{LJ}(r_{ij};\lambda) = U_{LJ}[r_{ij}^{LJ}(\lambda;\alpha)] \tag{2.7}$$

and

$$U_{SC}^{Coul}(r_{ij};\lambda) = U_{LJ}[r_{ij}^{Coul}(\lambda;\beta)] \tag{2.8}$$

where, α and β are adjustable positive semidefinite parameters for the *LJ* and *Coul* softcore interactions, respectively, with values of zero corresponding to no softcore modification for λ value. The long-range interaction cut-off, temperature, and pressure were set as described above for Molecular Dynamics simulations. During TI, the potentials were calculated "on the fly", i.e., as the simulation evolved through the different values of the λ parameter that progressively removed the non-bonded interactions of the ligand (i.e., the TS) using softcore potentials to transition from a bound to an unbound state. For each system (i.e., WT CRISPR-Cas12a, and the N1009A, F1012A, K1013A, R1014A, R1016A, K1018A mutants), we considered two states A and B based on the reaction coordinate values along the minimum free energy path of TS traversal (Figure 3.4a). Alchemical simulations were performed at 21 values of λ (from 0 to 1, in steps of 0.05), reaching $\sim$30 ns for each window. Convergence of the $\Delta\Delta G$ for each system is reported in Supplementary Figure B.16. Considering all the systems and states investigated here, a total of $\sim$8.8 μs sampling was reached.

B.7 FnCas12a Expression and Purification

The DNA sequence of *Francisella tularensis* subsp. novicida U112 (Fn)Cas12a (WP_003040289) was codon optimized for heterologous expression in *Escherichia coli* (E. coli) and synthesized by GeneArt (Thermo Fisher Scientific). The FnCas12a gene was inserted into the 1B plasmid (Addgene #29653) using ligation-independent cloning (LIC), resulting in a construct carrying an N-terminal hexahistidine tag followed by a tobacco etch virus (TEV) protease cleavage site. Point mutations were introduced by Gibson assembly of the PCR-amplified vector backbone with gBlock Gene Fragments (IDT) encoding the individual mutations. The sequences of the synthetic gene, primer and gBlocks are listed in Supplementary Table B.4. Mutant FnCas12a constructs were purified as for wild type. Purification of Cas12a was done as described.[145,146] In brief, Cas12a constructs were expressed in *E. coli* BL21 Rosetta2 (DE3) cells (Novagen, Wisconsin, USA). Cells were lysed in 20 mM Tris pH 8.0, 500 mM NaCl, 5 mM imidazole, 1 μg/mL pepstatin, 200 μg/mL 4-(2-Aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF) by ultrasonication. Clarified lysate was applied to a 10 ml Ni-NTA (Sigma-Aldrich) affinity column. The column was washed with 20 mM Tris pH 8.0, 500 mM NaCl, 5 mM imidazole, and bound protein was eluted by increasing imidazole concentration to 250 mM. Eluted protein was dialysed against 20 mM HEPES pH 7.5, 250 mM KCl, 1 mM DTT, 1 mM EDTA overnight at 4 °C in the presence of TEV protease to remove the 6xHis-affinity tag. Cleaved protein was further purified using a HiTrap HP Heparin column (GE Healthcare, Illinois, USA), eluting with a linear gradient to 1.0 M KCl. Elution fractions were pooled, concentrated, and further purified by size exclusion chromatography using a Superdex 200 (16/600) column (GE Healthcare) in

20 mM HEPES-KOH pH 7.5, 500 mM KCl, 1 mM DTT yielding pure, monodisperse proteins. Aliquots were flash-frozen in liquid nitrogen and stored at -80 °C.

B.8 FnCas12a Nuclease Activity Assays

In vitro nuclease activity assays were conducted using purified WT or mutant FnCas12a proteins programmed with a crRNA targeting the ⌊-sequence (oMS017, IDT, HPLC purified) and a dsDNA substrate containing fluorescently labeled TS (oDS285:oDS271, Merck, HPLC purified). FnCas12a and crRNA were mixed in a ratio of 1:1.2 and incubated for 10 min at 25 °C to allow binary complex formation. Reactions were started by addition of target dsDNA (FnCas12a:dsDNA, 10:1) and incubated at 37 °C. All samples were assembled in a final reaction volume of 20 μL containing 0.5 μM (mutant) FnCas12a, 0.6 μM crRNA and 50 nM dsDNA in a final buffer of 10 mM HEPES-KOH pH 7.5, 250 mM KCl, 5 mM MgCl2, 0.5 mM DTT. Reactions were stopped at indicated time points by addition of EDTA and Proteinase K (Thermo Fisher Scientific) in final concentrations of 80 mM and 0.8 mg/mL, respectively and incubated for 15 min at 37 °C. Samples were mixed with equal volume of a 2X dPAGE loading dye (95 % formamide, 25 mM EDTA), heated to 95 °C for 5 min and resolved on a 15 % denaturing (7 M Urea) polyacrylamide gels run in 0.5 X TBE buffer. Assays were conducted in technical triplicates; indicated error bars represent the standard error of mean. Fluorescence of the ATTO532-labeled substrate and cleavage products was detected using a Typhoon FLA 9500 gel imager; the cleavage rate was quantified based

on loss of uncleaved substrate DNA using ImageQuant TL v.8.2.0. Sequences of RNA and DNA oligos utilized in the nuclease assays are listed in Table B.3.

B.9 Supplementary Tables

Table B.1: **Description of the simulated complexes.**

Sl. No.	**Name**	**Salient Features**
1	Cryo-EM Cas12a	PDB ID: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) (Stella et al. *Cell* **2018**, 165, 1856-1871). Contains α-helical lid (L1008 – K1021).
2	X-ray Cas12a	PDB ID: 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) (Swarts et al. Mol. *Cell* **2017**, 66, 221-233). Contains an unstructured loop form of the lid as reconstructed using Homology Modelling (details in the Supplementary Methods).

Continued on next page

continued from previous page

Sl. No.	**Name**	**Salient Features**
3	Complete Cas12a	Based on PDB ID: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation) and including a longer TS downstream of the crRNA:TS duplex, taking cues from PDB ID: 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop) (details in the Supplementary Methods).
4	Catalytically Competent Cas12a	Based on the complete Cas12a system and PDB: 5U30 [https://doi.org/10.2210/pdb5U30/pdb] (AacC2c1-sgRNA-extended target DNA ternary complex) (Yang et al. *Cell* **2016**, 167, 1814-1828). The 5'-tail of the TS downstream of the crRNA:TS duplex, which includes the scissile phosphate, locates in the RuvC active (details in the Supplementary Methods).

Table B.2: **Details for the calculation of bending angles using vectors.** For each bending angle, we selected two regions of interest (mask 1 and mask 2), used as a reference to compute the vectors ($\vec{V}$) for angle analysis. Details are in the Supplementary Methods.

Bending angle	Mask 1	Mask 2
crRNA:TS arching	crRNA (A_{16}):TS (T_5)	crRNA (G_{19}):TS (C_2)
TS bending	Nucleobases C_2 – G_{-1}	Nucleobases G_{-1} – G_{-3}
Rec2 bending	Residues I370 – K382	Residues L451 – D470
Nuc bending	Residues H1187 – C1196	Residues F1240 – M1250

Table B.3: **Sequences of crRNA guide and target DNA substrates.**

ID	Description	Sequence (5' - 3')
oMS017	FnCas12a-crRNA ⌊	AAUUUCUACUGUUGUAGAUGUGAUAAGUGGAAUGCCAUGUGGG
oDS271	NTS ⌊	ACGTGGAAGGATGCCTTTAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGT
oDS285	labeled TS ⌊	(ATTO532-)ACTCAATTTTGACAGCCCACATGGCATTCCACTTATCACTAAAGGCATCCTTCCACGT

Table B.4: **Sequences of DNA oligos, synthetic gene and gene fragments (gBlocks).**

ID	Description	Sequence
oMS413	Backbone amplification primer fwd	CAGCTCCGTTTGAGACCTTTAAAAAG
oMS414	Backbone amplification primer rev	CCAGTCTTTCCTTGCGCTATCC

Continued on next page

continued from previous page

ID	Description	Sequence
FnCas12a	Codon optimized gene (*E. coli*)	ATGGCAAGCATTTATCAGGAATTTGTTAATAAATATAGCCTGAGCAAAACCTTA CGTTTTGAGCTGATACCGCAAGGTAAAACATTAGAAAACATTAAAGCTCGTGGT CTCATTCTGGATGATGAGAAACGTGCAAAAGACTACAAAAAGGCAAAACAGATT ATTGATAAATATCATCAATTTTTTATTGAGGAGATTCTGAGCTCGGTTTGTATT TCGGAGGATCTGCTGCAAAACTATTCTGATGTTTATTTTAAATTAAAAAAGAGC GATGATGATAATTTACAGAAAGATTTTAAAAGCGCTAAAGATACTATTAAGAAA CAGATTTCTGAATATATTAAGGACAGCGAGAAATTTAAGAATCTCTTTAATCAG AACTTAATAGATGCAAAAAAAGGGCAGGAGAGCGATCTGATTTTATGGTTAAAG CAGTCTAAGGATAATGGTATTGAATTATTTAAAGCTAATAGCGATATAACAGAT ATTGATGAGGCTCTGGAAATTATAAAATCTTTTAAAGGTTGGACAACCTATTTT AAGGGTTTTCATGAAAATCGTAAAAATGTTTATAGCTCGAATGATATTCCGACA TCTATTATTTATAGGATTGTCGATGATAATCTCCCGAAATTTTTAGAAAATAAA GCAAAGTATGAGAGCCTGAAAGACAAAGCACCGGAAGCAATTAACTATGAACAG ATTAAAAAAGATTTGGCTGAAGAGTTAACTTTTGATATTGACTACAAAACATCT

Continued on next page

continued from previous page

ID	Description	Sequence
		GAAGTTAATCAGCGTGTTTTTAGCTTAGATGAAGTTTTTGAGATTGCTAACTTT AATAATTATTTAAATCAGAGCGGTATTACCAAATTTAATACCATTATTGGTGGT AAATTTGTCAATGGTGAAAATACAAAGCGTAAAGGTATTAATGAATATATTAAT TTATACAGCCAACAGATTAATGATAAAACACTAAAAAAATATAAAATGAGCGTT CTGTTTAAGCAGATTCTGAGCGATACAGAATCTAAATCTTTTGTCATTGATAAG CTGGAAGATGATAGCGATGTCGTTACAACTATGCAGAGCTTTTATGAGCAGATT GCTGCATTTAAAACAGTCGAAGAAAAATCTATTAAAGAAACATTATCTCTGCTG TTTGATGATCTGAAAGCACAGAAATTAGATTTGAGCAAAATTTATTTTAAAAAT GATAAATCTTTAACCGATTTAAGCCAGCAGGTTTTTGATGATTATAGCGTTATT GGTACAGCTGTCTTAGAATATATTACCCAGCAGATTGCTCCGAAAAATTTAGAT AACCCGAGCAAGAAAGAGCAGGAACTGATTGCTAAAAAAACCGAAAAAGCTAAA TACCTGTCTTTAGAAACCATTAAGTTAGCTCTGGAAGAATTTAATAAGCATCGT GATATTGATAAACAATGTAGGTTTGAAGAAATTTTAGCTAACTTTGCTGCAATT CCCATGATTTTTGATGAAATTGCACAGAACAAAGACAATCTCGCTCAAATTTCT

Continued on next page

continued from previous page

ID	Description	Sequence
		ATAAAATATCAGAATCAGGGTAAAAAAGACTTATTACAGGCAAGCGCTGAAGAT GATGTTAAAGCAATAAAGGATTTACTGGATCAGACCAATAATCTACTGCATAAA TTAAAAATTTTTCATATTAGCCAAAGCGAAGATAAGGCTAATATTCTGGACAAG GATGAGCATTTTTATTTAGTCTTTGAGGAGTGCTACTTTGAGTTAGCTAATATT GTACCGTTATATAACAAAATTCGTAACTATATTACCCAGAAGCCGTATAGCGAT GAGAAATTTAAGCTAAATTTTGAGAACTCGACCCTCGCAAATGGTTGGGATAAA AATAAAGAGCCGGACAATACTGCTATTCTGTTTATAAAAGATGATAAATATTAT CTAGGTGTAATGAATAAGAAAAATAACAAAATTTTTGATGATAAAGCAATAAAA GAAAATAAAGGGGAGGGTTATAAAAAAATTGTTTATAAATTACTGCCGGGGGCT AATAAAATGCTGCCGAAGGTTTTCTTTTCTGCAAAATCTATTAAATTTTATAAT CCGAGCGAAGATATTTTACGTATTCGTAATCATTCGACACATACAAAAAATGGT AGCCCGCAGAAAGGGTATGAAAAATTTGAGTTTAATATTGAAGATTGCAGGAAA TTTATTGATTTTTATAAACAATCTATTAGCAAGCATCCCGAGTGGAAAGATTTT GGGTTTCGTTTTTCTGATACCCAGCGTTATAATTCTATTGATGAATTTTATCGT

Continued on next page

continued from previous page

ID	Description	Sequence
		GAAGTTGAAAATCAGGGGTACAAATTAACCTTTGAAAATATTAGCGAGTCGTAT ATTGATTCGGTCGTTAATCAAGGTAAACTCTACTTATTCCAGATATATAATAAA GATTTTAGCGCATATTCGAAAGGGAGGCCGAATTTACATACCCTGTATTGGAAA GCTCTATTTGATGAGCGTAATTTACAGGATGTAGTTTATAAGTTAAATGGTGAG GCTGAGTTATTTTATAGAAAACAGAGCATTCCGAAAAAAATAACCCACCCGGCA AAAGAGGCTATTGCAAATAAAAACAAAGATAATCCGAAAAAAGAGAGCGTTTTT GAATATGATCTGATAAAAGATAAAAGGTTTACCGAAGATAAGTTTTTCTTTCAC TGTCCGATTACAATAAATTTTAAATCTAGCGGGGCAAATAAGTTTAATGATGAA ATAAATCTGCTCTTAAAAGAAAAAGCTAATGATGTTCATATTCTGAGCATTGAT CGTGGTGAACGTCATCTGGCATACTATACCCTCGTCGATGGTAAAGGGAATATA ATAAAACAGGATACCTTCAACATAATTGGTAATGATCGTATGAAAACAAACTAC CATGATAAGTTAGCAGCTATTGAGAAAGATAGGGATAGCGCAAGGAAAGACTGG AAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGTATTTATCTCAAGTCGTT CATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAATTGTAGTTTTTGAGGAT

Continued on next page

continued from previous page

ID	Description	Sequence
		CTGAATTTTGGGTTTAAACGTGGGAGATTCAAGGTCGAGAAGCAAGTATATCAG AAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATTTAGTTTTCAAAGATAAT GAGTTTGATAAAACCGGGGGGGTATTACGTGCATATCAATTAACAGCTCCGTTT GAGACCTTTAAAAAGATGGGTAAACAGACAGGTATTATATACTATGTCCCGGCA GGTTTTACCAGCAAAATTTGTCCGGTCACCGGTTTTGTCAATCAACTGTATCCG AAGTATGAAAGCGTATCGAAATCTCAGGAGTTCTTTAGCAAGTTTGACAAGATT TGTTATAACTTAGATAAGGGGTATTTTGAGTTTAGCTTTGATTATAAAAACTTT GGTGACAAGGCAGCTAAAGGGAAGTGGACCATTGCATCGTTTGGGAGCCGTCTC ATTAACTTTCGTAATAGCGATAAAAATCATAATTGGGATACCAGGGAAGTTTAT CCGACCAAAGAGTTGGAGAAACTCTTAAAAGATTATTCTATAGAATATGGGCAT GGGGAATGTATAAAAGCTGCAATTTGCGGTGAGTCGGACAAAAAGTTTTTTGCA AAGTTAACCAGCGTATTAAATACCATACTGCAAATGAGAAACAGCAAAACAGGT ACCGAGCTGGATTATTTAATTAGCCCGGTCGCTGATGTCAATGGGAATTTCTTT GATTCGAGGCAAGCTCCGAAAAATATGCCGCAGGATGCAGATGCTAATGGTGCA TATCATATTGGGTTAAAAGGTCTAATGTTATTAGGTAGGATAAAAAATAATCAG GAGGGGAAAAAACTAAATCTCGTTATAAAAAATGAAGAGTATTTTGAGTTCGTA CAAAATAGGAATAACTAA

Continued on next page

continued from previous page

ID	Description	Sequence
gMS061	gBlock for K1013A mutation	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGTTTGCACGTGGGAGATTCAAGGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC
gMS062	gBlock for R1014A mutation	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGTTTAAAGCAGGGAGATTCAAGGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC
gMS063	gBlock for R1016A mutation	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGTTTAAACGTGGGGCATTCAAGGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC

Continued on next page

continued from previous page

ID	Description	Sequence
gMS064	gBlock for R1018A mutation	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGTTTAAACGTGGGAGATTCGCAGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC
gMS065	gBlock for E1006Q mutation	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTCAAGATCTGAATTTTGGGTTTAAACGTGGGAGATTCAAGGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC
gMS066	gBlock for K1013A/R1014A mutations	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGTTTGCAGCAGGGAGATTCAAGGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC

Continued on next page

continued from previous page

ID	Description	Sequence
gMS067	gBlock for R1014A /K1018A mutations	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGTTTAAAGCAGGGAGATTCGCAGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC
gMS068	gBlock for K1013A/ R1014A/ R1016A/ K1018A mutations	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGTTTGCAGCAGGGGCATTCGCAGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC
gMS069	gBlock for N1009A mutation	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGGCATTTGGGTTTAAACGTGGGAGATTCAAGGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC

Continued on next page

continued from previous page

ID	Description	Sequence
gMS070	gBlock for F1012A mutation	GCGCAAGGAAAGACTGGAAAAAGATTAATAACATAAAAGAGATGAAAGAGGGGT ATTTATCTCAAGTCGTTCATGAAATTGCAAAGTTAGTTATTGAGTATAATGCAA TTGTAGTTTTTGAGGATCTGAATTTTGGGGCAAAACGTGGGAGATTCAAGGTCG AGAAGCAAGTATATCAGAAGCTGGAAAAAATGTTAATTGAGAAATTAAACTATT TAGTTTTCAAAGATAATGAGTTTGATAAAACCGGGGGGGGTATTACGTGCATATC AATTAACAGCTCCGTTTGAGACC

Table B.5: **DNA Cleavage Assay Time Points Data.**

	N1009A	**F1012A**	**K1013A**	**R1014A**	**R1016A**	**K1018A**	**K1013A, R1014A, R1016A, K1018A**	**F1012A, K1013A, R1014A, R1016A, K1018A**	**E1006Q**
0 min									
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SD	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SEM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1 min									
	-8.681	-46.786	-21.200	-3.234	-17.130	-4.165	-45.206	-45.709	-49.617
	-9.363	-46.530	-19.954	-3.653	-16.042	-2.050	-44.844	-45.076	-49.326
	-10.236	-47.728	-18.670	-2.448	-16.913	-2.461	-44.899	-47.280	-49.067

Continued on next page

continued from previous page

	N1009A	**F1012A**	**K1013A**	**R1014A**	**R1016A**	**K1018A**	**K1013A, R1014A, R1016A, K1018A**	**F1012A, K1013A, R1014A, R1016A, K1018A**	**E1006Q**
Mean	-9.427	-47.015	-19.941	-3.112	-16.695	-2.892	-44.983	-46.022	-49.337
SD	0.780	0.631	1.265	0.612	0.576	1.121	0.195	1.135	0.275
SEM	0.450	0.364	0.730	0.353	0.332	0.647	0.112	0.655	0.159
5 min									
	-2.397	-44.561	-15.250	-2.294	-15.081	-4.660	-38.223	-44.048	-75.958
	-2.099	-43.431	-14.485	-1.557	-14.379	-2.696	-37.651	-43.673	-76.756
	-3.034	-44.506	-13.778	-0.078	-15.342	-3.819	-36.246	-44.618	-77.141
Mean	-2.510	-44.166	-14.504	-1.310	-14.934	-3.725	-37.373	-44.113	-76.618
SD	0.478	0.637	0.736	1.128	0.498	0.985	1.017	0.476	0.603
SEM	0.276	0.368	0.425	0.651	0.287	0.569	0.587	0.275	0.348
15 min									

Continued on next page

continued from previous page

	N1009A	**F1012A**	**K1013A**	**R1014A**	**R1016A**	**K1018A**	**K1013A, R1014A, R1016A, K1018A**	**F1012A, K1013A, R1014A, R1016A, K1018A**	**E1006Q**
	-1.567	-31.064	-9.530	-0.615	-11.246	-2.524	-29.005	-33.559	-83.333
	-0.416	-31.257	-9.648	0.500	-10.296	-1.698	-29.335	-32.736	-84.463
	-1.063	-33.638	-9.751	0.612	-11.524	-1.876	-28.215	-35.379	-85.603
Mean	-1.015	-31.986	-9.643	0.166	-11.022	-2.033	-28.852	-33.891	-84.466
SD	0.577	1.434	0.111	0.678	0.644	0.435	0.576	1.352	1.135
SEM	0.333	0.828	0.064	0.392	0.372	0.251	0.332	0.781	0.655
60 min									
	-0.944	-19.340	-5.185	-0.480	-7.744	-1.686	-20.268	-25.082	-86.198
	0.153	-20.069	-5.429	0.549	-6.350	-0.664	-19.624	-25.421	-88.457
	-0.281	-19.726	-6.354	0.104	-6.995	-1.401	-20.940	-27.032	-87.997
Mean	-0.357	-19.712	-5.656	0.058	-7.029	-1.250	-20.277	-25.845	-87.550

Continued on next page

continued from previous page

	N1009A	**F1012A**	**K1013A**	**R1014A**	**R1016A**	**K1018A**	**K1013A, R1014A, R1016A, K1018A**	**F1012A, K1013A, R1014A, R1016A, K1018A**	**E1006Q**
SD	0.553	0.365	0.617	0.516	0.698	0.528	0.658	1.042	1.194
SEM	0.319	0.210	0.356	0.298	0.403	0.305	0.380	0.601	0.689

B.10 Supplementary Figures

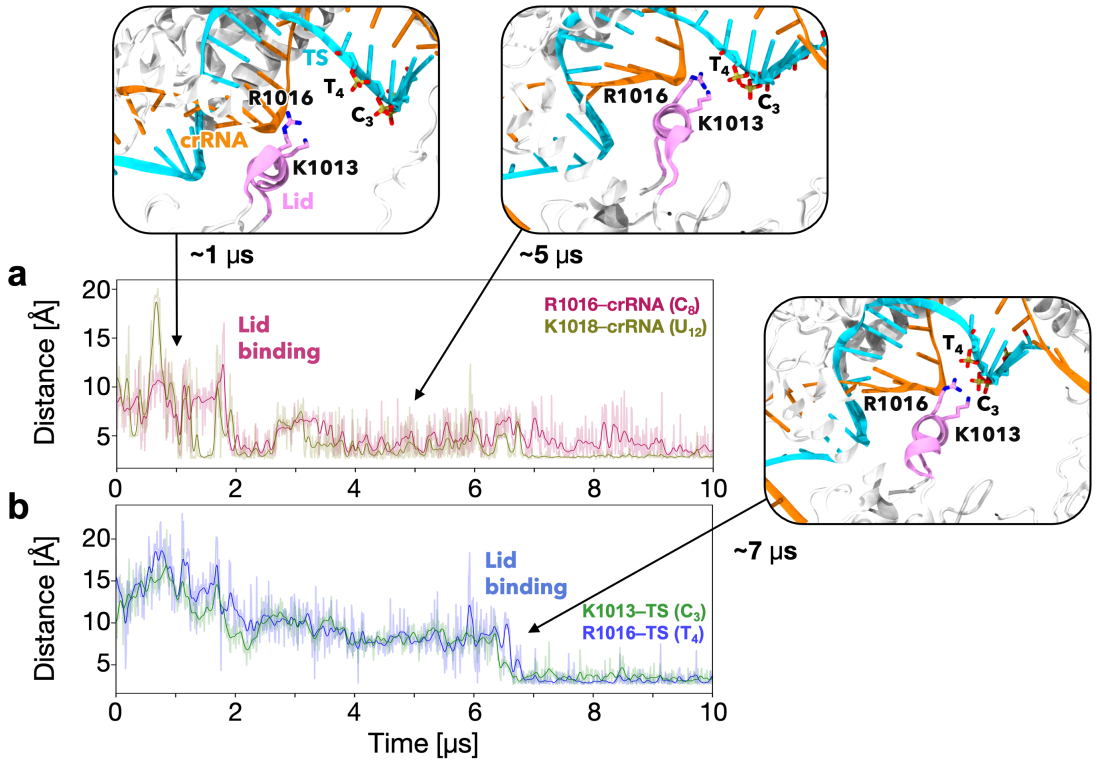


Figure B.1: **Molecular dynamics (MD) simulations of the 6GTG[1] cryo-EM structure of CRISPR-Cas12a (Figure 3.1a).** Interactions established by the positively charged residues of the α-helical lid with **(a)** the crRNA (R1016:NH1 – C_8:OP1 and K1018:NZ – U_{12}:OP1) and **(b)** the DNA TS (K1013:NZ – C_3:OP1, R1016:NH2 – T_4:OP1) evolving over ∼10 μs of molecular dynamics (MD) simulations. Interactions are measured as oxygen-nitrogen distance and are considered formed under 4 Å distance.[116] Representative snapshots from MD simulations are indicated by the arrows, showing the progressively increasing interaction between the α-helical lid (mauve) and the crRNA:TS duplex (crRNA: orange, TS: cyan; top panel and side panel). The R1016 and K1018 residues form intermittent interactions with the crRNA backbone at positions 8 and 12 respectively, especially after ∼4 μs of MD simulations. K1013 and R1016 approach the DNA TS at positions 3 and 4 respectively and form stable electrostatic interactions upon ∼7 μs of MD (i.e., persistent interactions through the remainder of the MD run). The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation.

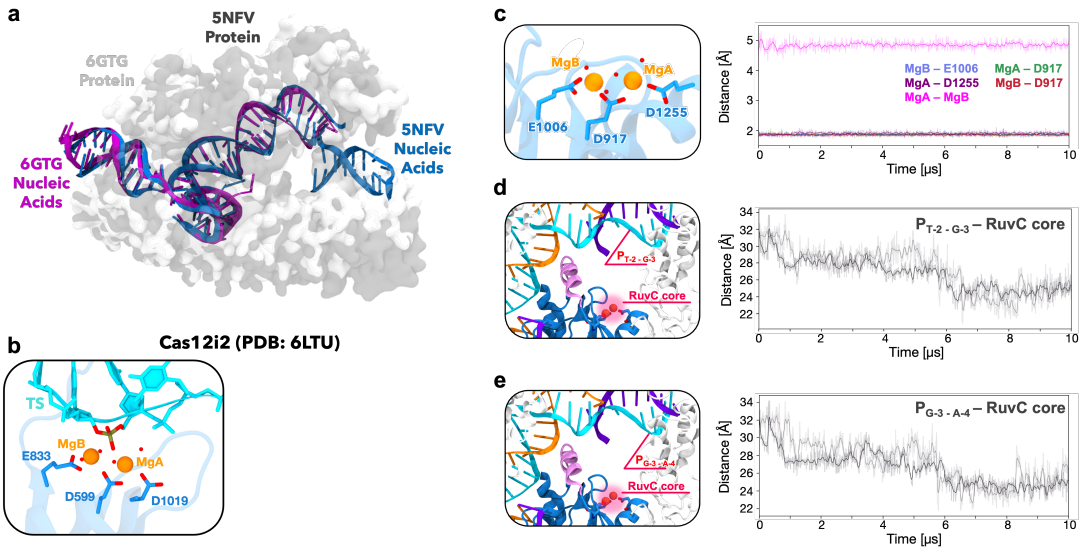


Figure B.2: **Structure and dynamics of the complete CRISPR-Cas12a system (Figure 3.1c).** **(a)** Superimposition of the cryo-EM structure (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation)) onto the X-ray structure (PDB: 5NFV [https://doi.org/10.2210/pdb5NFV/pdb][2] (FnCas12a bound to R-loop)) of CRISPR-Cas12a. The protein is shown in molecular surface (white for 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) and gray for 5NFV [https://doi.org/10.2210/pdb5NFV/pdb][2] (FnCas12a bound to R-loop)) and the nucleic acids are shown in ribbons (purple for 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) and blue for 5NFV [https://doi.org/10.2210/pdb5NFV/pdb][2] (FnCas12a bound to R-loop)). More details are in the Supplementary Methods. **(b)** X-ray structure of CRISPR-Cas12i2 (PDB: 6LTU [https://doi.org/10.2210/pdb6LTU/pdb][169] (Cas12i2 ternary complex)), capturing the catalytic site of the RuvC domain in the presence of the DNA TS substrate interacting with the catalytic Mg_{2+} ions. Relevant distances of the catalytic architecture are enlisted in the table. **(c)** Distance analyses between the key catalytic residues (D917, E1006 and D1255) and the two catalytic Mg_{2+} ions (MgA and MgB) evolving over $\sim$10 μs MD simulation, exhibiting noteworthy stability of the catalytic core. A representative snapshot of the RuvC catalytic core with the catalytic Mg_{2+} ions is shown with the graph. **(d - e)** Distance between the centre of mass of the RuvC catalytic core (left panel, composed of D917, E1006 and D1255 residues coordinating with two Mg_{2+} ions) and the scissile phosphate (P_{SCI}), i.e., **(d)** between T_{-2} and G_{-3}, and **(e)** G_{-3} and A_{-4}, as also shown with representative snapshots. The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation.

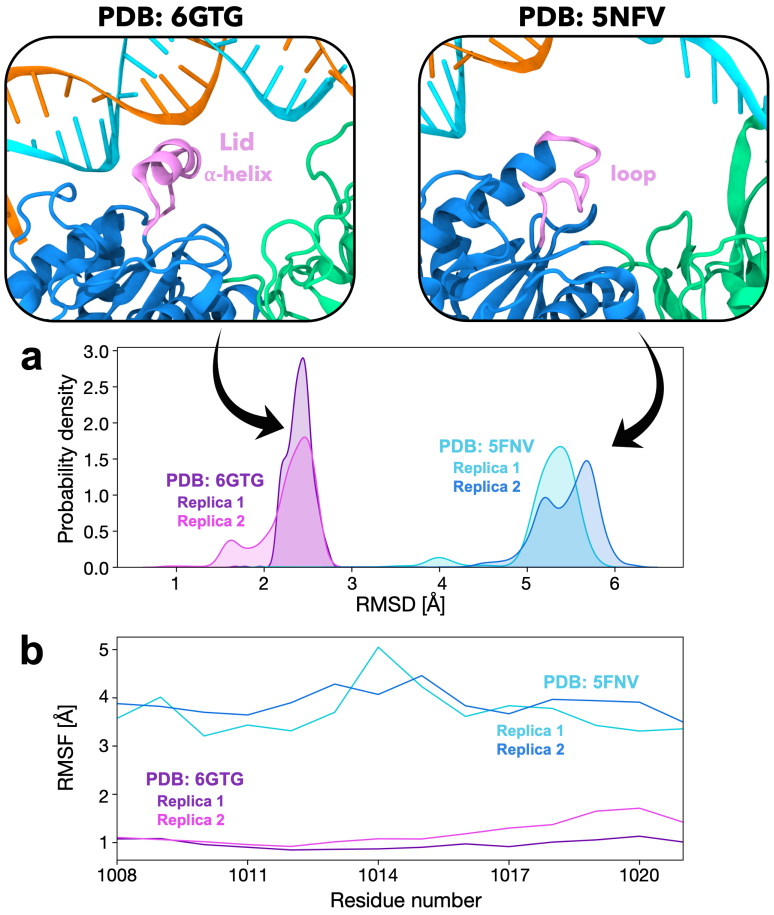


Figure B.3: **Stability (a) and flexibility (b) of the lid (residues L1008–K1021) along MD simulations.** **(a)** Probability density plots of the all-atom root-mean-square deviation (RMSD) of the lid along MD simulations of the cryo-EM structure PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) of CRISPR-Cas12a including a complete TS (magenta) and of the X-ray structure PDB: 5NFV [https://doi.org/10.2210/pdb5NFV/pdb][2] (FnCas12a bound to R-loop) (blue), with respect to the structures obtained experimentally. Hydrogen atoms were excluded from the RMSD calculation. The 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation) displays an α-helical lid (top panel, left), while the X-ray structure 5NFV [https://doi.org/10.2210/pdb5NFV/pdb][2] (FnCas12a bound to R-loop) reports an unstructured loop (top panel, right). Data are reported for two simulation replicates of $\sim$10 μs each. **(b)** Root-mean-square fluctuations (RMSF) of the Cα atoms of the lid along MD simulations of the CRISPR-Cas12a systems defined above. Markedly higher RMSD and RMSF are observed for the lid assuming the shape of an unstructured loop as compared to the stable α-helical lid.

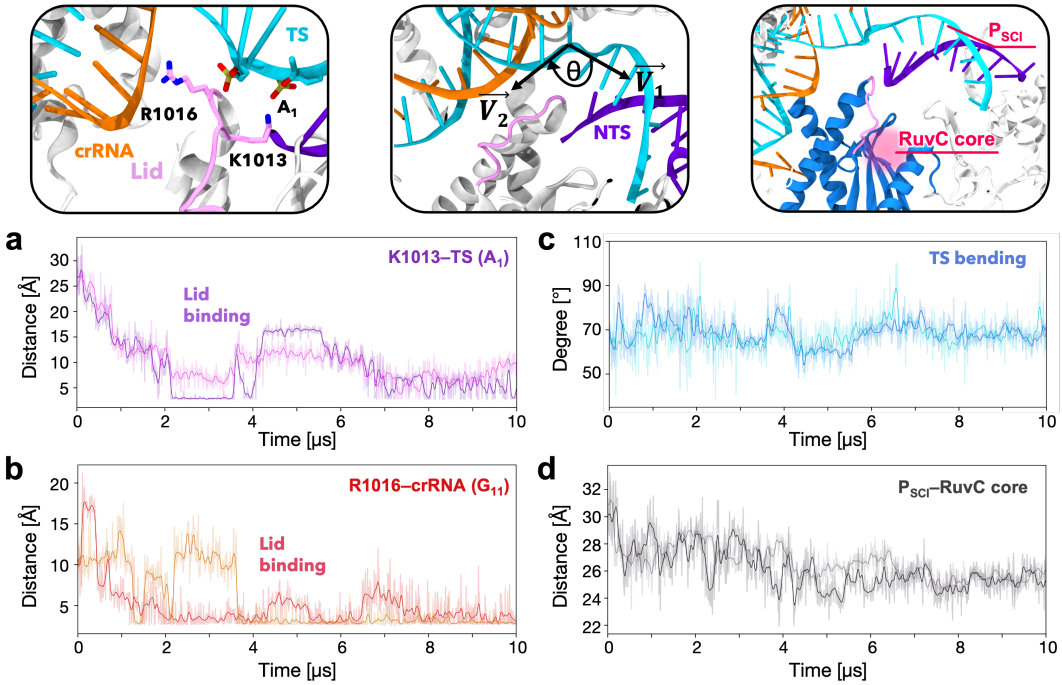


Figure B.4: **Molecular dynamics (MD) simulations of the 5NFV[2] X-ray structure of CRISPR-Cas12a (Figure 3.1b).** Molecular simulations were extended to $\sim$10 μs in two replicates. **(a - b)** Time evolution of the distances between positively charged residues of the unstructured lid and the crRNA:TS duplex (top panel, left). The top panel **(a)** reports the interaction between the lid and the TS (K1013:NZ – A_1:OP1), while the bottom panel **(b)** shows the interaction between the lid and the crRNA (R1016:NH1 – G_{11}:OP1). Interactions are measured as oxygen-nitrogen distance and are considered formed under 4 Å distance.[116] **(c)** Angle between the vectors used to represent the bending of the TS, as shown in the cartoon representation (top panel, middle) and explained in the Supplementary Methods. **(d)** Distance between the scissile phosphate $(P)_{SCI}$ between positions -2 and -3) in the DNA target strand (TS) and the RuvC catalytic pocket (D1255, D917, and E1006; top panel, right). The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation.

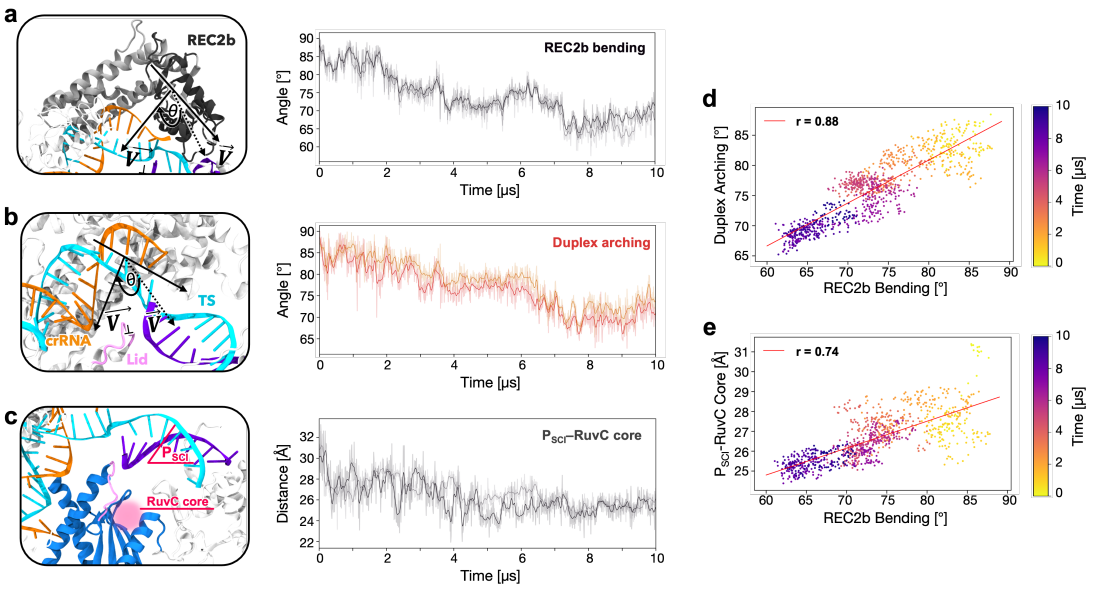


Figure B.5: **Bending of major domains and effects on the scissile phosphate along MD simulations of the 5NFV[2] X-ray structure of CRISPR-Cas12a (Figure 3.1b).** **(a - b)** Time evolution of the bending angles describing **(a)** the bending of REC2b; **(b)** the arching of the terminal major groove of the crRNA:TS duplex. Angles between the vector passing through the region of interest and its perpendicular out of the plane were analysed with respect to the first frame of MD production runs. A representative snapshot of the system is reported on the left of each graph, showing the analysed angles (details in the Supplementary Methods). **(c)** Distance between the scissile phosphate (P_{SCI} between positions -2 and -3) in the DNA target strand (TS) and the RuvC catalytic pocket. The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation. **(d - e)** Scatter plots between **(d)** REC2b bending and duplex arching (a – b), where Pearson correlation coefficient, r = 0.88; and **(e)** REC2b bending and P_{SCI} – RuvC core distance (a, c), where Pearson correlation coefficient, r = 0.74. Data are reported for two simulation replicates of $\sim$10 μs each and the correlation coefficient was calculated combining data from two replicates.

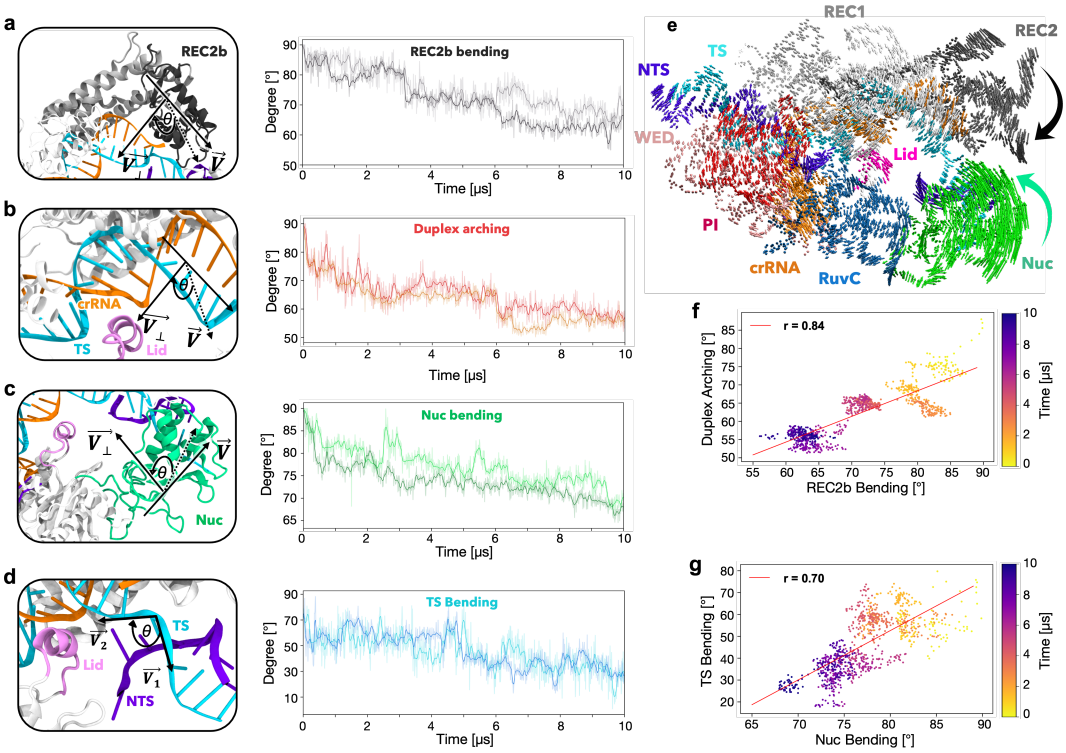


Figure B.6: **Bending of major domains along MD simulations of the complete CRISPR-Cas12a system (Figure 3.1c). (a - d)** Time evolution along MD simulations of bending angles describing **(a)** the bending of REC2b; **(b)** the arching of the terminal major groove of the crRNA:TS duplex; **(c)** the bending of Nuc and **(d)** the bending of the DNA TS region formed by bases G_{-1}, G_1, C_2, C_3 and T_4. Angles between the vector passing through the region of interest and its perpendicular out of the plane were analysed with respect to the first frame of MD production runs. A representative snapshot of the system is reported on the left of each graph, showing the analysed angles (details in the Supplementary Methods). The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation. **(e)** "Essential dynamics", derived from the first principal component (PC1) of the complete CRISPR-Cas12a system, shown using arrows of sizes proportional to the amplitude of motions. **(f - g)** Scatter plots between **(f)** REC2b bending and duplex arching (a – b), where Pearson correlation coefficient, r = 0.84; and **(g)** Nuc bending and TS bending (c – d), where Pearson correlation coefficient, r = 0.70. Data are reported for two simulation replicates of $\sim$10 μs each and the correlation coefficient was calculated combining data from two replicates.

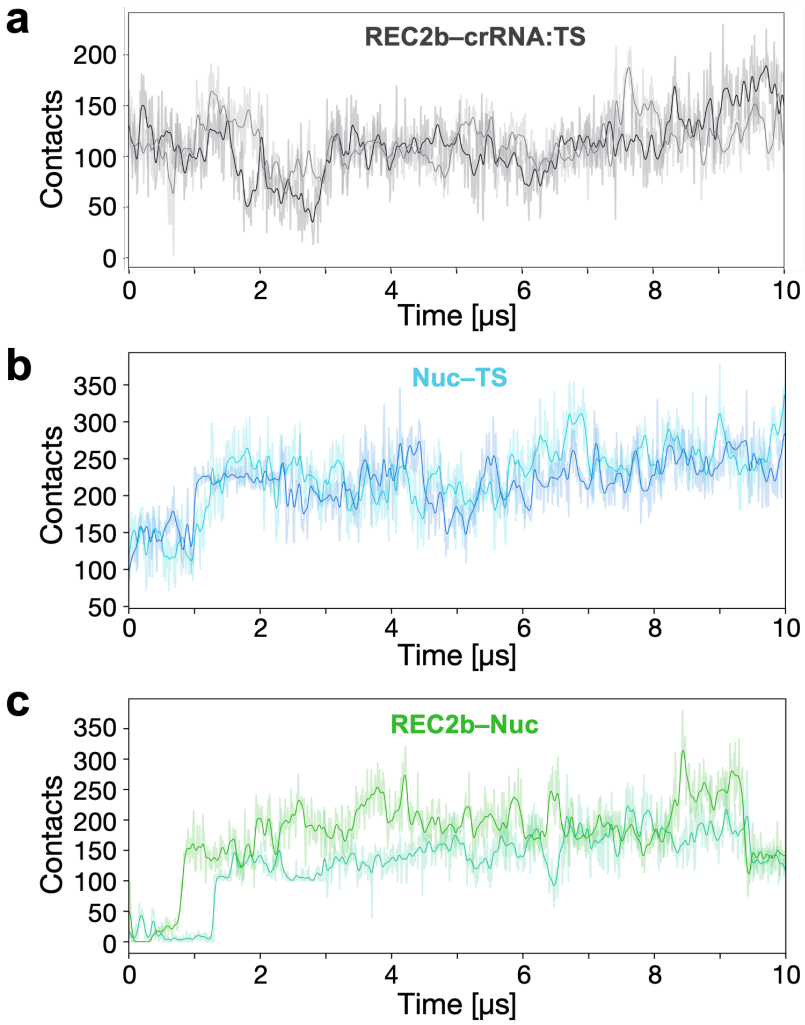


Figure B.7: **Number of contacts between major domains of Cas12a and nucleic acids.** Molecular dynamics simulations of the complete CRISPR-Cas12a system show a significant number of contacts between: **(a)** the REC2b region and the crRNA:TS duplex; **(b)** the Nuc domain and the DNA target strand (TS); and **(c)** Rec2b and Nuc. Data are reported for two simulation replicates of $\sim$10 μs each. A contact is considered when the distance between two heavy atoms among the regions of interest is less than 3.5 Å. The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation.

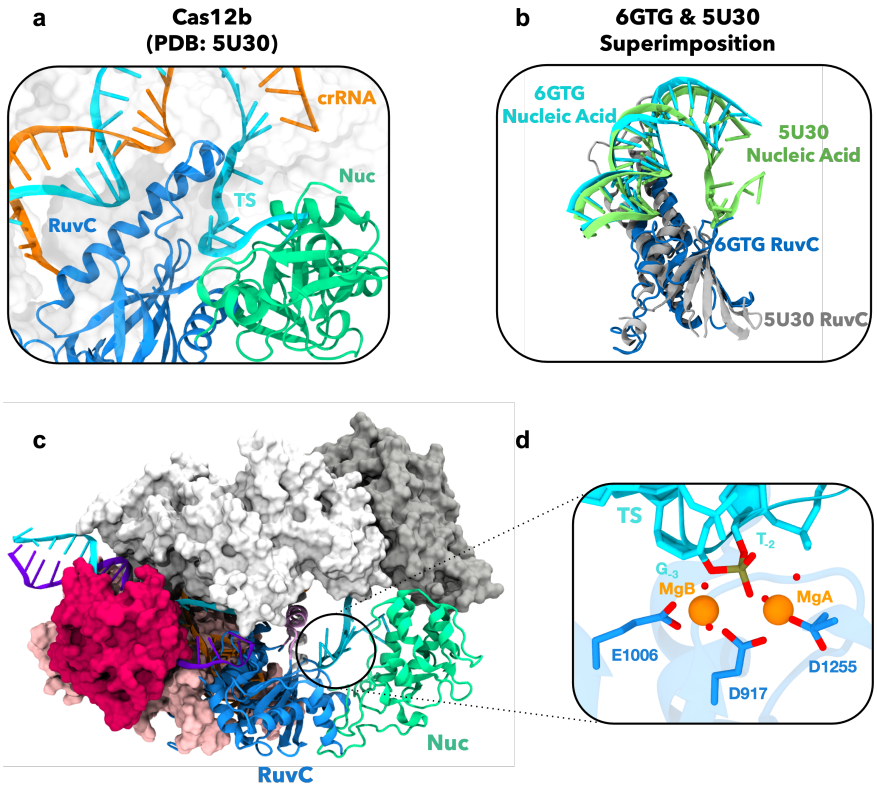


Figure B.8: **Overview of the catalytically competent CRISPR-Cas12a built from related X-ray structures.** The protein is shown in molecular surface, highlighting the RuvC (blue) and Nuc (green) domains as ribbons. The crRNA (orange), the DNA target strand (TS, cyan) and non-target strand (NTS, violet) are also shown as ribbons. Mg^{2+} ions are shown as spheres. **(a)** X-ray structure of CRISPR-Cas12b (PDB: 5U30 [https://doi.org/10.2210/pdb5U30/pdb][172] (AacC2c1-sgRNA-extended target DNA ternary complex)), showing the accommodation of the TS at the RuvC catalytic cleft. The TS bends down from the crRNA:TS duplex toward the active site, forming a U-shaped strand with its 5'-tail interacting with the Nuc domain. **(b)** Superimposition of the RuvC region and part of the nucleic acid from the cryo-EM structure (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation)) of Cas12a (protein in blue and nucleic acids in cyan) and the X-ray structure (PDB: 5U30 [https://doi.org/10.2210/pdb5U30/pdb][172] (AacC2c1-sgRNA-extended target DNA ternary complex)) of Cas12b (protein in gray and nucleic acid in gray) sharing structural homology. c Catalytically competent Cas12a for TS cleavage, including the TS in the RuvC catalytic cleft. The model was built using the structures in **(a)** and **(b)** as references (details are reported in the SI text). **(d)** Close-up view of the RuvC active site, accommodating the TS in the two-metal ion architecture.

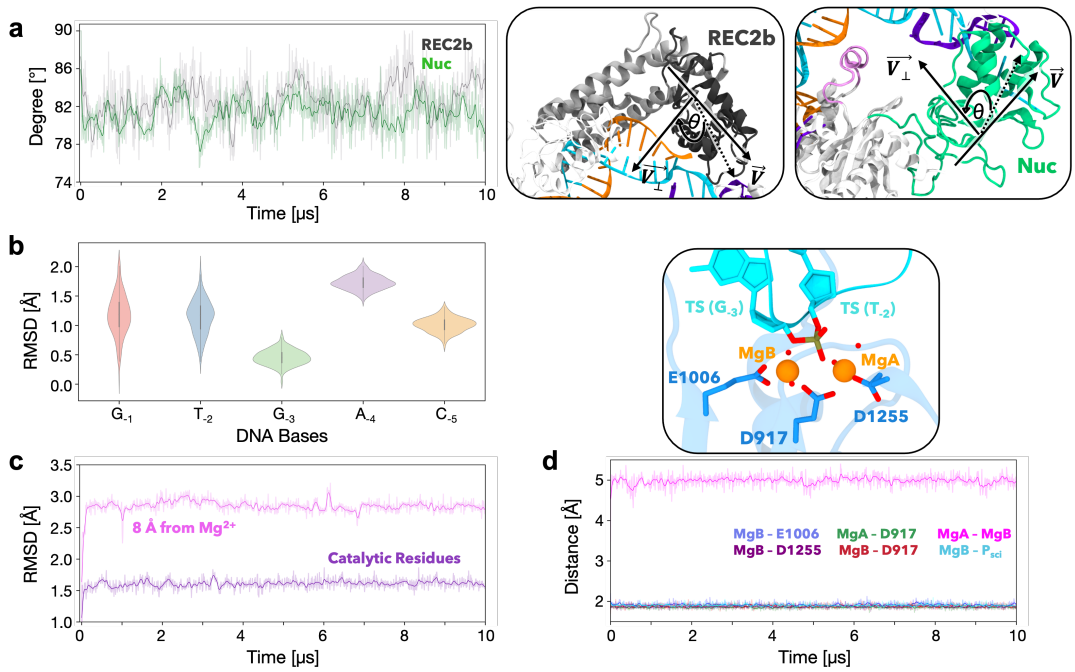


Figure B.9: **Molecular dynamics simulations of the catalytically competent CRISPR-Cas12a for TS cleavage.** **(a)** Time evolution of the angles describing the bending of REC2b (gray) and Nuc (green) along $\sim$10 μs long MD simulations. Angles between the vector passing through the region of interest and its perpendicular out of the plane were analysed with respect to the first frame of MD production runs. A representative snapshot on the right panel shows the analysed angles (details in the Supplementary Methods). This analysis shows that REC2b and Nuc assume a stable conformation. **(b - d)** Stability of the RuvC active site. **(b)** All-atom root-mean-square deviation (RMSD) of the DNA TS nucleobases close to the catalytic core. The RMSD of the nucleobases interacting with the catalytic Mg_{2+} ions involves T_{-2} (T_{-2}:OP1 – MgA; RMSD $<$ 1.5 Å) and G_{-3} (G_{-3}:OP3 – MgB; RMSD $<$ 0.5 Å), report remarkable stability. The flanking nucleobases also show appreciably low RMSD, highlighting the stability of the TS at the active site. **(c)** Time evolution of the RMSD of the catalytic residues around 8 Å of Mg_{2+} ions (pink), and the D917, E1006 and D1255 residues (violet). This analysis reveals that the RuvC catalytic site with the DNA TS is remarkably stable along $\sim$10 μs of MD simulations. **(d)** Distance analyses between the key catalytic residues (D917, E1006 and D1255) and the two catalytic Mg_{2+} ions (MgA and MgB) evolving over $\sim$10 μs MD simulation, exhibiting noteworthy stability of the catalytic core. A representative snapshot of the RuvC catalytic core with the DNA TS is shown above the graph. The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation.

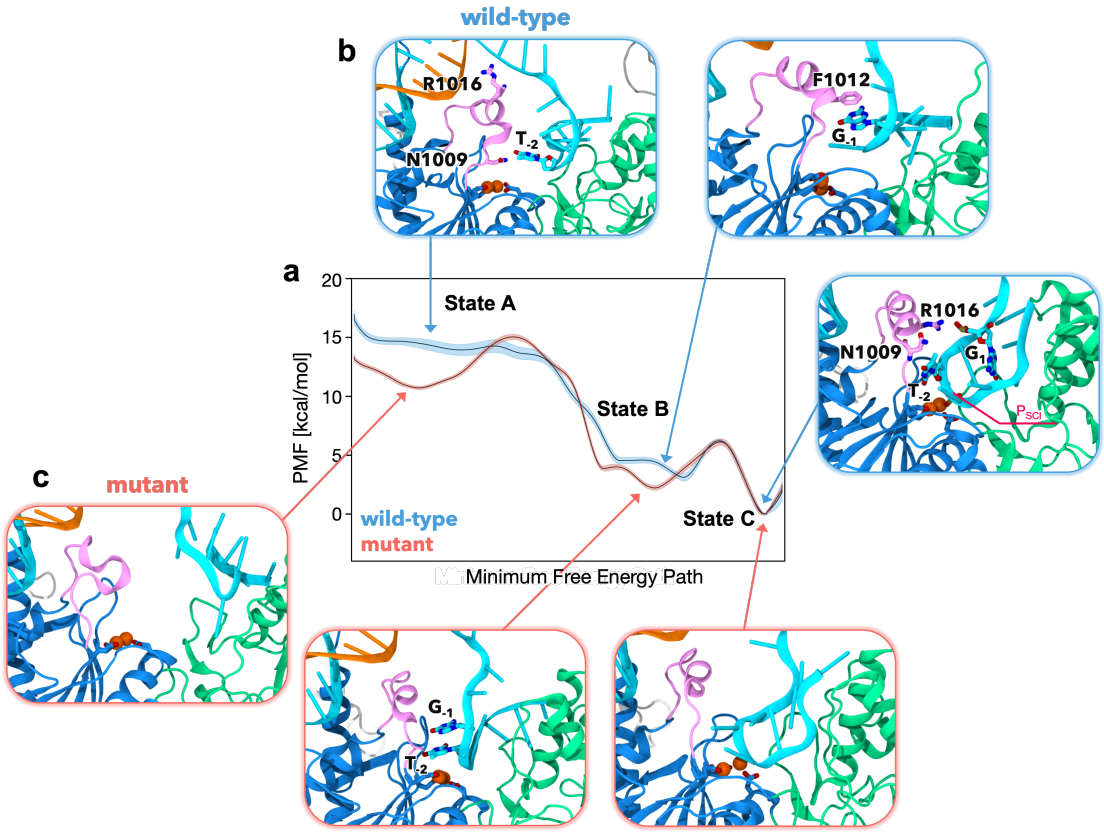


Figure B.10: **Free energy profiles for the traversal of the DNA target strand (TS) toward the RuvC catalytic core.** **(a)** Free energy profiles for the wild-type (WT) Cas12a (blue) and upon mutating relevant residues of the α-helical lid (F1012, N1009, K1013, R1014, R1016, K1018) into alanine (red). The Potential of Mean Force (PMF) is computed from 2-D umbrella sampling simulations and plotted along the minimum free energy path (see the Supplementary Methods). Representative snapshots, indicated by arrows, are based on the reaction coordinate values along the minimum free energy path of the wild-type Cas12a **(b)** and its mutant (c).

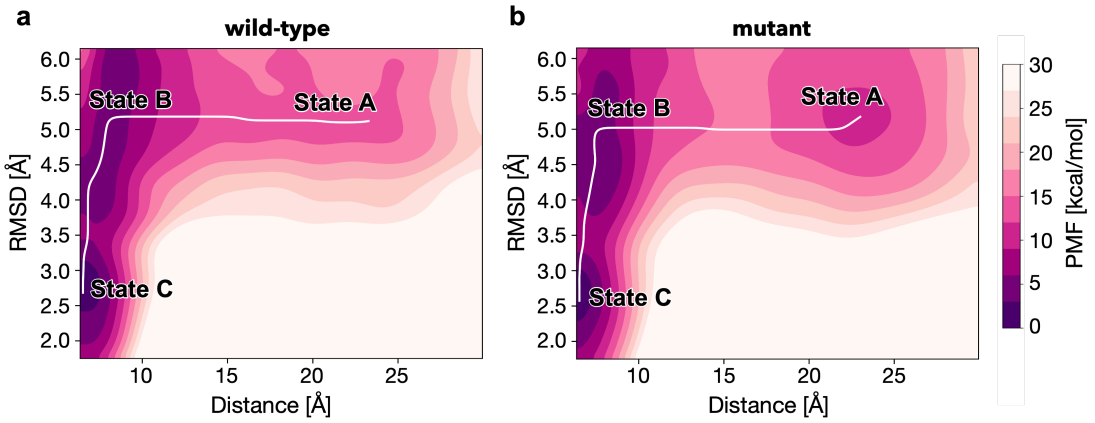


Figure B.11: **Two-dimensional (2-D) free energy maps for the traversal of the DNA target strand (TS) toward the RuvC core. (a - b)** 2-D Umbrella Sampling simulations for the wild-type Cas12a **(a)** and upon mutating critical residues of the α-helical lid (F1012, N1009, K1013, R1014, R1016, K1018) into alanine (b). The Potential of Mean Force (PMF) is computed along two Reaction Coordinates (RCs). RC1 is the root-mean-square deviation (RMSD) of the Cα atoms between the initial and the final state. RC2 is the distance between the centre of mass of the RuvC catalytic core (D917, E1006, D1255) and the TS region including the scissile phosphate (i.e., the DNA nucleobases at positions -2 to -4, as indicated by biochemical and single-molecule experiments). Full details are reported in the Supplementary Methods. The solid white line represents the minimum free energy path connecting the initial and final states. The different states along the minimum free energy path are indicated (also in Supplementary Figure B.8).

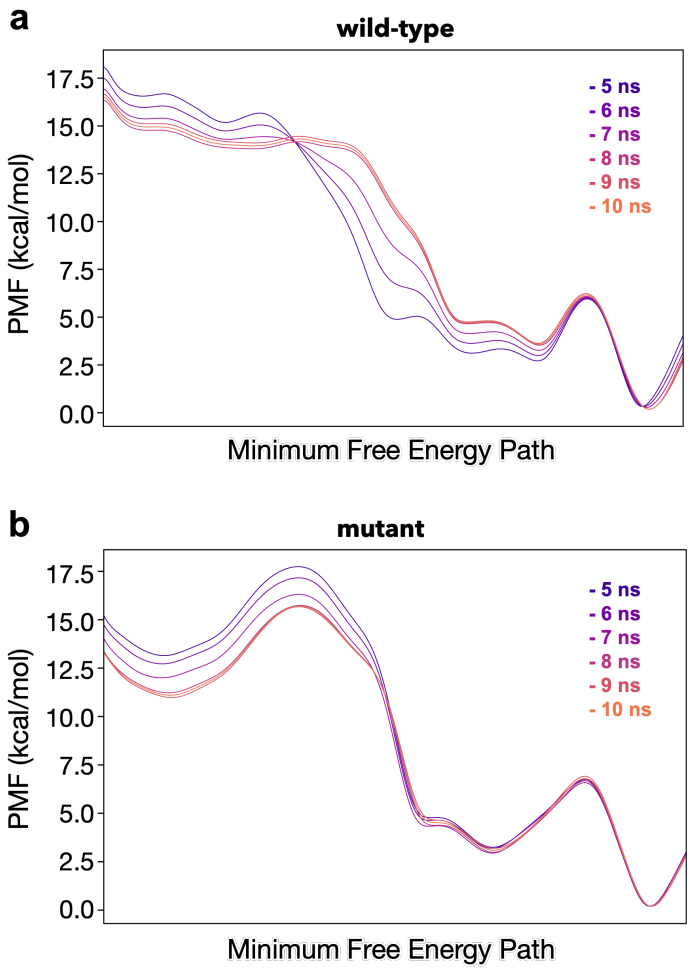


Figure B.12: **Convergence of the free energy profiles obtained through Umbrella Sampling (US) simulations. (a - b)** Convergence of the potential of mean force (PMF) along the minimum free energy path for US simulations of the wild-type CRISPR-Cas12a (a), and including alanine mutations of critical residues of the α-helical lid (b). N1009, F1012, K1013, R1014, R1016 and K1018 were mutated into alanine to probe their role in the traversal of the DNA target strand. The PMF (kcal/mol) was computed for sampling until 5 ns, 6 ns, 7 ns, 8 ns, 9 ns and 10 ns after removing the biases incurred by the restraints during the sampling. Convergence is reached ∼8 ns onward for both the systems and we used bin sizes of ∼10 ns to compute our 2-D free energy surfaces (Supplementary Figure B.9) and the minimum free energy paths (Supplementary Figure B.10 and Figure 3.4).

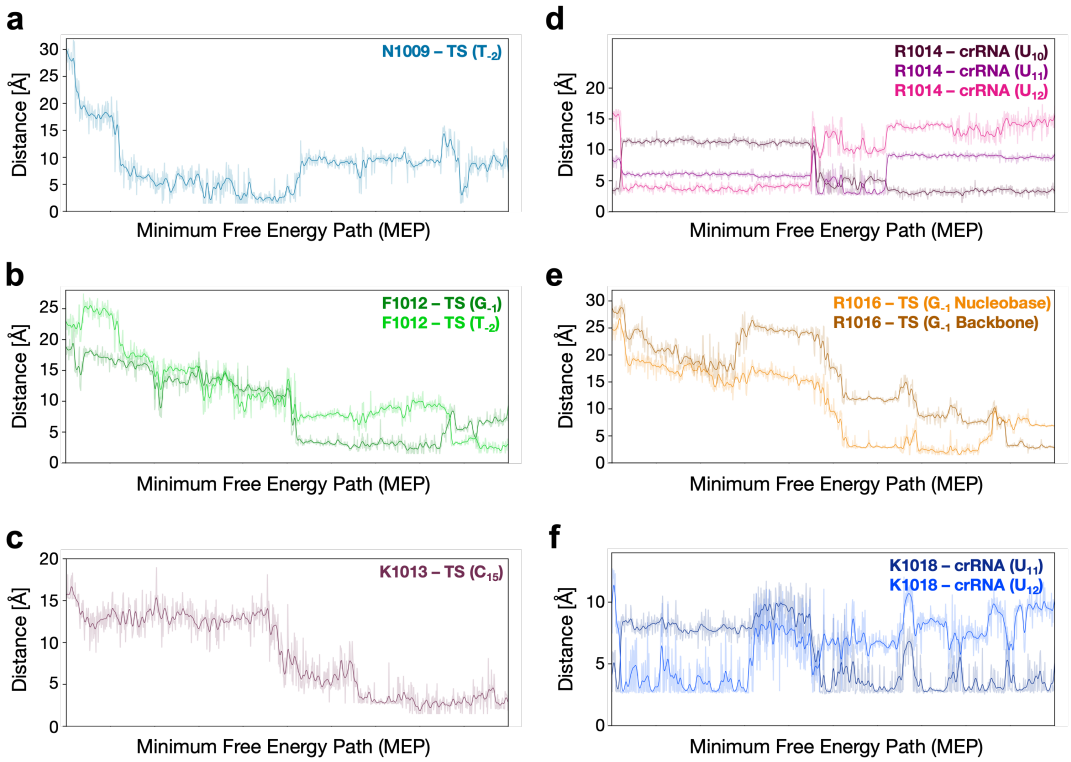


Figure B.13: **Interactions between the lid and the crRNA:TS duplex along the minimum energy path (MEP) from Umbrella Sampling simulations of the wild-type CRISPR-Cas12a.** **(a - f)** Distances between **(a)** N1009 – TS (N1009:ND2 – T_{-2}:OP1), **(b)** F1012 – TS (F1012:phenylalanine ring – G_{-1}:imidazole ring and F1012:phenylalanine ring – T_{-2}:pyrimidine ring), **(c)** K1013 – TS (K1013:KZ – C_{15}:OP1), **(d)** R1014 – crRNA (R1014:NH2 – U_{10}:OP1, R1014:NH2 – U_{11}:OP1 and R1014:NH2 – U_{12}:OP1), **(e)** R1016 – TS (R1016:NH2 – G_{-1}:nucleobase and G_{-1}:OP1), and **(f)** K1018 – crRNA (K1018:NZ – U_{11}:OP1 and K1018:NZ – U_{12}:OP1). The translucent lines show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation. The free energy profiles along the MEP are reported in Figure 3.4a and Supplementary Figure B.10.

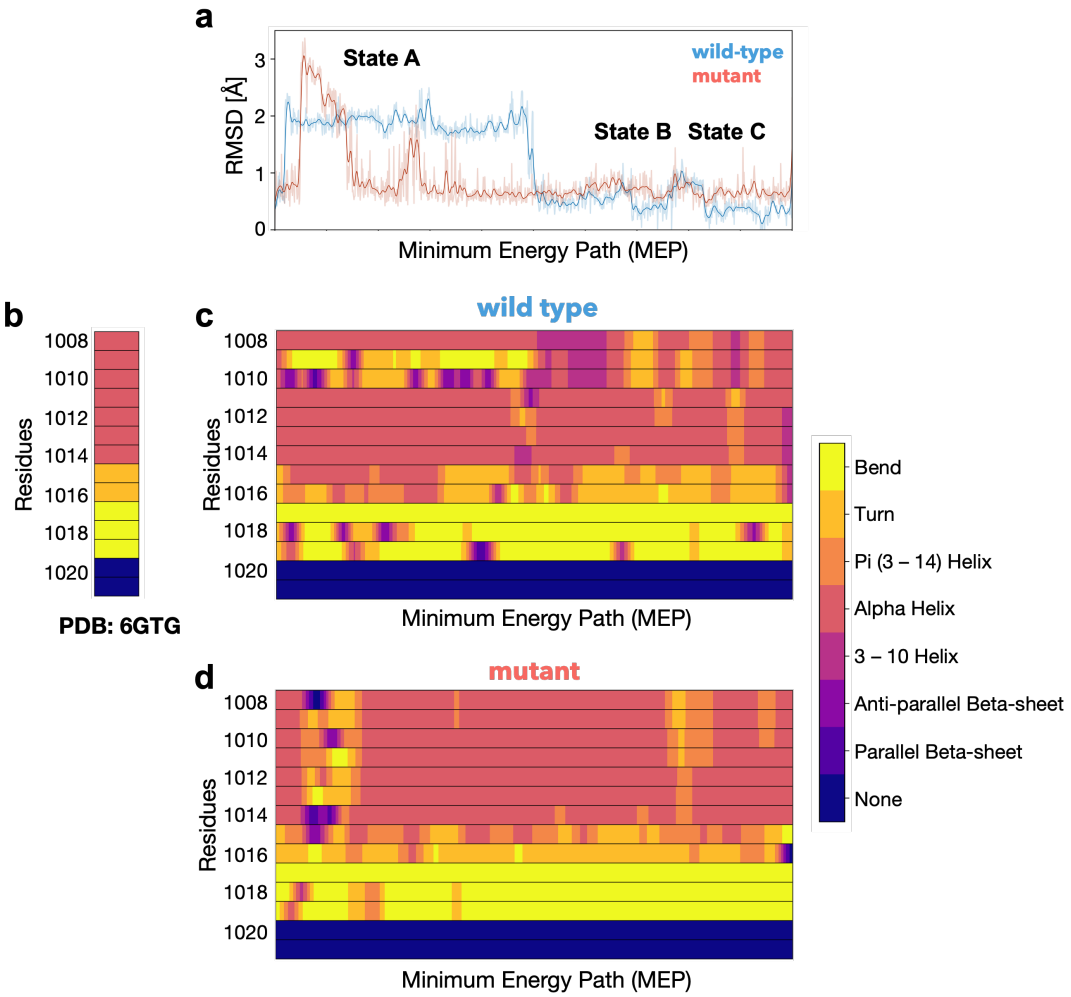


Figure B.14: **Lid dynamics along the minimum energy path (MEP) of the Umbrella sampling simulation.** a Root-mean-square deviation (RMSD) of the lid (L1008 – K1021) along the minimum free energy path of the Umbrella sampling simulations (Figure 3.4 and Supplementary Figure B.10) for the wild-type (blue) and mutant (red) systems. The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation. **(b - d)** Secondary structure analysis of the lid residues using the DSSP algorithm[300,301] for **(b)** the cryo-EM Cas12a structure (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb][1] (Cas12a – I4 Conformation)), **(c)** wild-type, and **(d)** mutant systems. The secondary structure analysis for the wild-type and mutant systems were performed along the minimum free energy path of the Umbrella sampling simulations (Supplementary Figures B.10 and B.11).

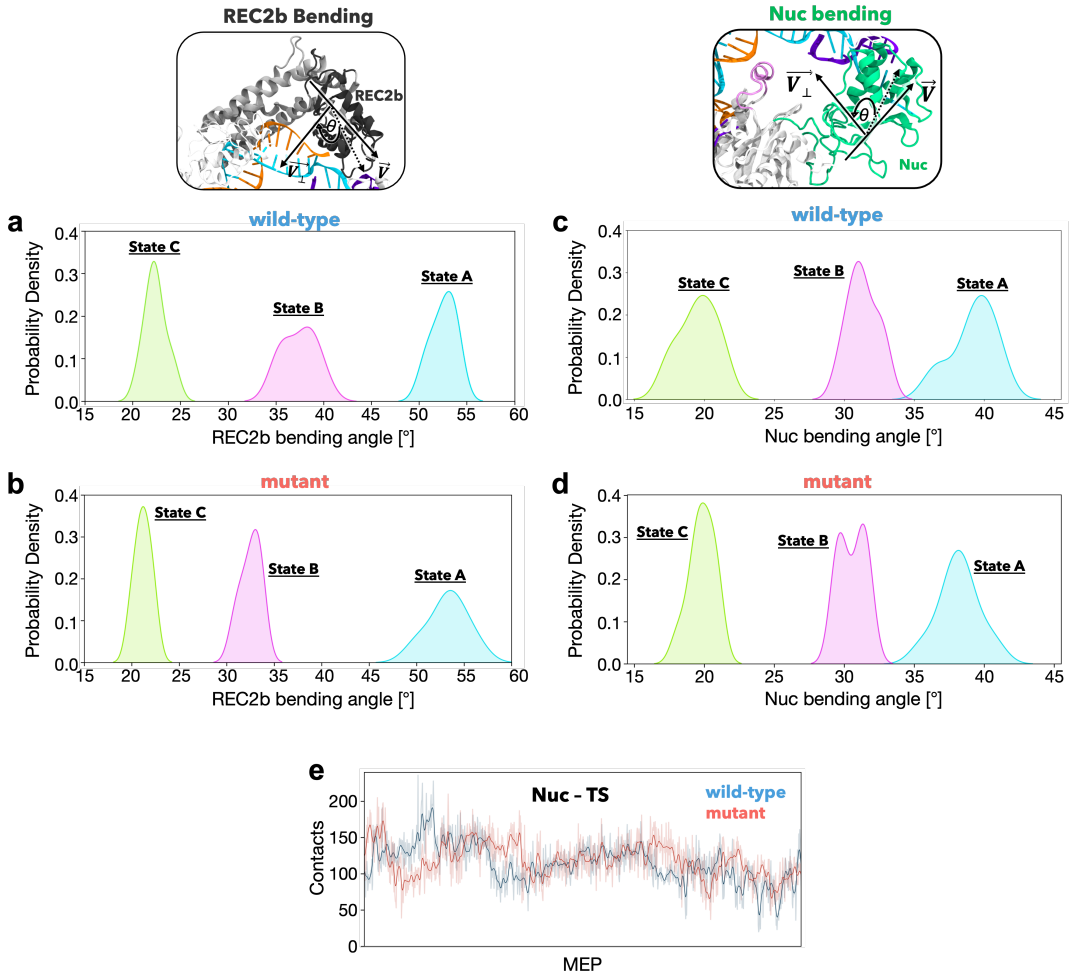


Figure B.15: **Dynamics of major domains at different states along the free energy profiles of the Umbrella sampling simulations. (a - d)** Bending angle of the REC2b region and Nuc domain is computed at three different states (states A-C; Supplementary Figures B.10 and B.11) along the minimum energy path (MEP) of the umbrella sampling simulations. Bending of (a – b) REC2b and (c – d) Nuc of the wild-type Cas12a and its mutant. Moving from state A to C, REC2b and Nuc moves inward, as indicated by a decrease of its bending angle in the representative snapshots. **(e)** Contacts between Nuc and TS along the MEP. A contact is considered when the distance between two heavy atoms among the regions of interest is less than 3.5 Å. The translucent lines in the plots show the actual data points, while the solid lines show the running average over 5 ns windows of MD simulation.

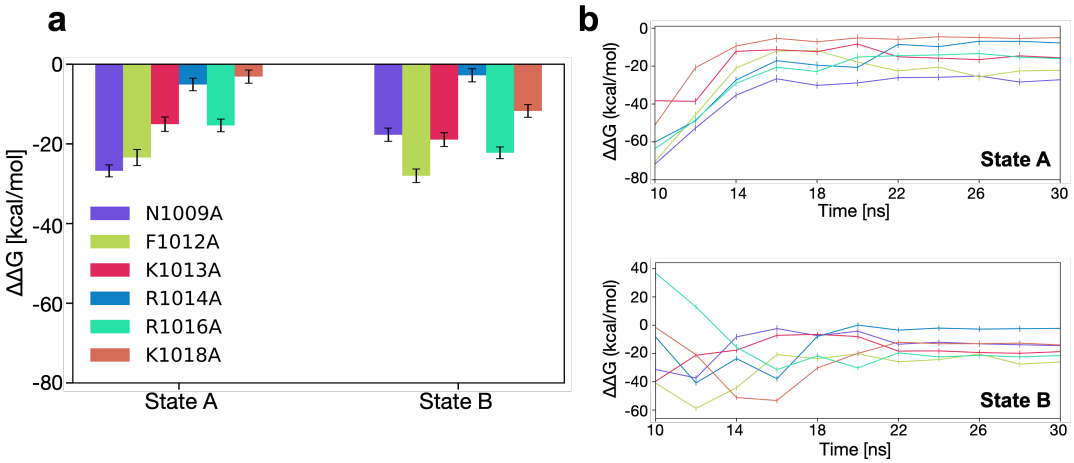


Figure B.16: **Convergence of the relative free energy changes obtained through alchemical free energy simulations (AFE).** a Relative free energy changes ($\Delta\Delta G$, in kcal/mol) of the individual residues with respect to alanine in binding the DNA TS at both the states encompassing the traversal of the TS guided by the α-helical lid, i.e., states A and B. Binding free energies are computed through the Multistate Bennett Acceptance Ratio (MBAR) method[118,119] (see the Methods section) and denoted with the associated MBAR error. **(b)** Convergence profile of the relative binding free energies computed until 10 ns, 12 ns, 14 ns, 16 ns, 18 ns, 20 ns, 22 ns, 24 ns, 26 ns, 28 ns and 30 ns of the alchemical free energy simulations are shown with the MBAR error. Convergence is reached ∼22 ns onward for both states A and B.

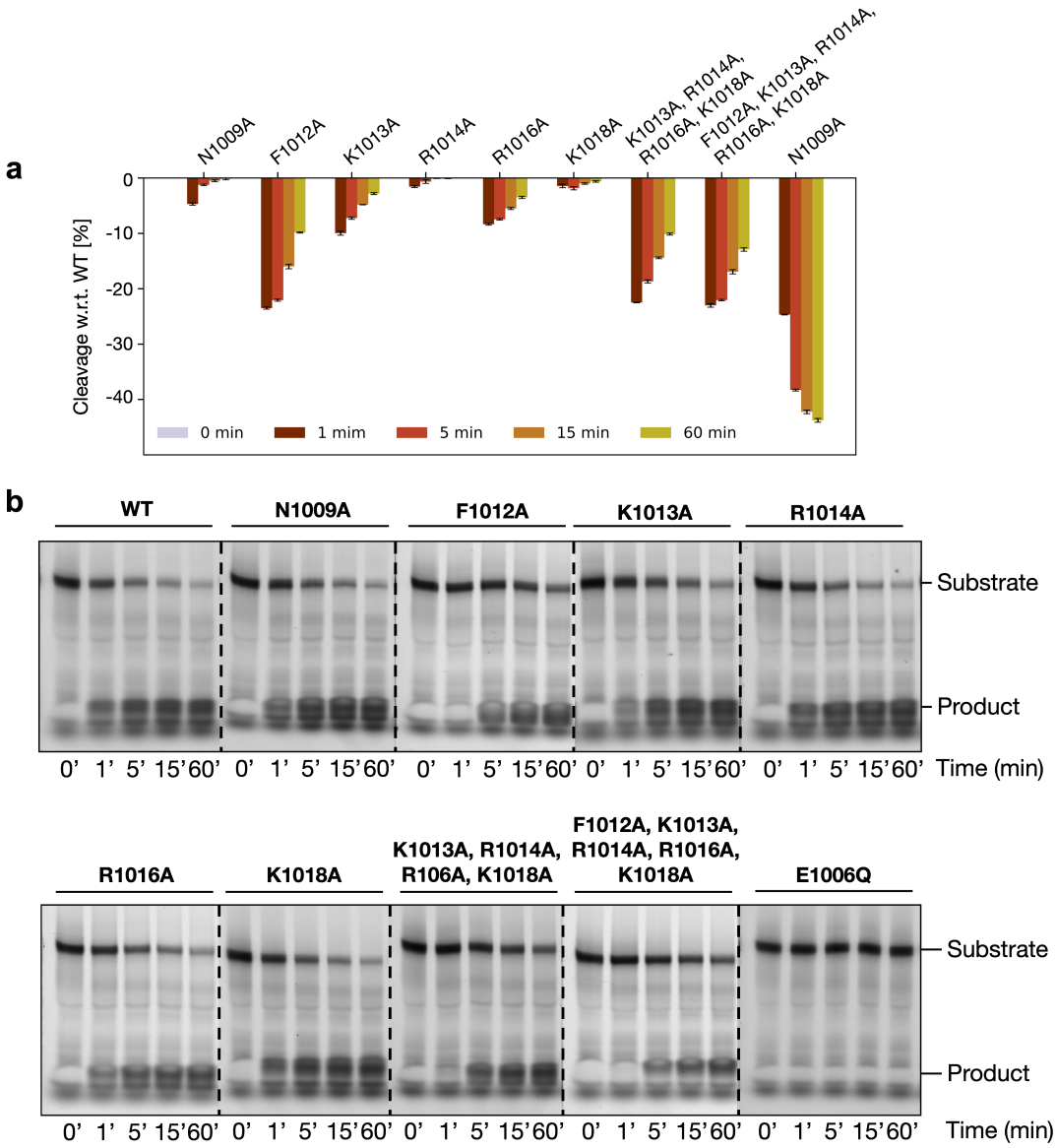


Figure B.17: ***In vitro* DNA TS cleavage assay of Cas12a mutants.** **(a)** DNA target strand (TS) cleavage percentage of Cas12a mutants compared to the wild type (WT). Data represents n = 3 independent replicates and the mean is plotted with the standard error of mean. **(b)** Denaturing polyacrylamide gels showing DNA TS cleaved products by Cas12a mutants at 0 min, 1 min, 5 min, 15 min and 60 min. Replica 1 from the three independent replicates has been shown here for each of the mutants (Supplementary Figure B.18, labelled with an asterisk sign).

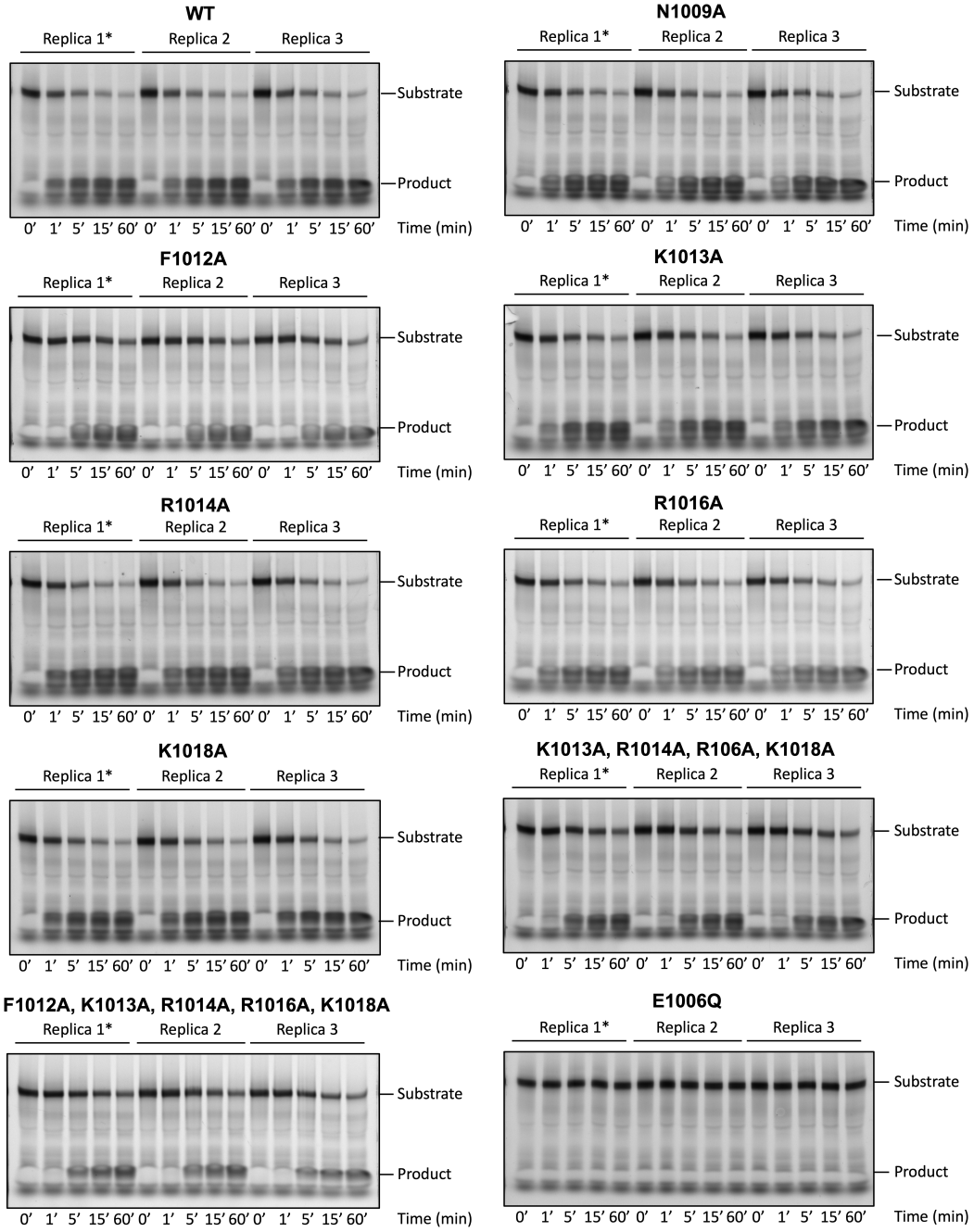


Figure B.18: ***In vitro*** **dsDNA cleavage assay of Cas12a mutants.** Cropped denaturing polyacrylamide gels showing DNA TS cleaved products by Cas12a mutants at 0 min, 1 min, 5 min, 15 min and 60 min. Data represents n = 3 independent replicates. Replica 1 has been labelled with an asterisk sign to indicate the replica shown in Supplementary Figure B.17.

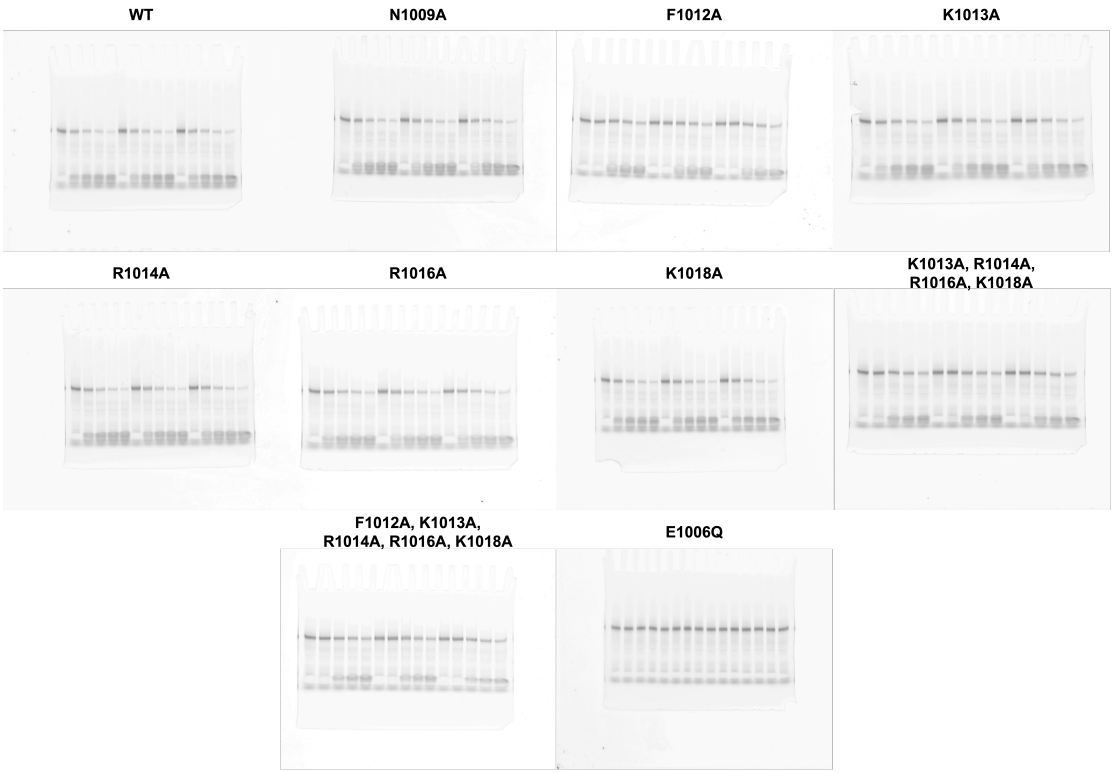


Figure B.19: ***In vitro*** **DNA TS cleavage assay of Cas12a mutants.** Uncropped images of denaturing polyacrylamide gels showing DNA TS cleaved products by Cas12a as used for Supplementary Figures B.17 and B.18.

B.11 Statistics and Reproducibility

All simulations were performed in replicates, as also specified in the main text and pertinent figure legends. Quantitation data of in vitro DNA cleavage assay represents single exponential fitting of n = 3 independent replicas and is plotted with the mean ± standard error of mean. Statistical significance was calculated using two-tailed unpaired t-test.

B.12 Ethics and Inclusion Statement

Molecular simulations and analyses were performed at the University of California, Riverside (United States), primarily by international graduate students and post-doctoral scholars. Experimental data were generated at the University of Zürich (Switzerland). We carefully evaluated and mutually agreed upon researcher contributions and authorship criteria, ensuring equity in research collaborations.

B.13 Data Availability Statement

All data are included in the article and Supplementary Information. Molecular simulations were based on two structures of the *F. novicida* Cas12a obtained upon NTS cleavage: (1) the cryo-EM structure EMD-0065 (PDB: 6GTG [https://doi.org/10.2210/pdb6GTG/pdb] (Cas12a – I4 Conformation)) solved at 3.27 Å resolution and (2) the X-ray structure PDB: 5NFV [https://doi.org/10.2210/pdb5NFV/pdb] (FnCas12a bound to R-loop), solved at 2.50 Å resolution. Sequences of oligos, synthetic genes and gene fragments are provided in Supplementary Table B.4. In vitro DNA cleavage assay time points data of Cas12a wild type and its mutants is provided in Supplementary Table B.5 and the corresponding uncropped gel images are shown in Supplementary Figure B.19. Atomic coordinates of the initial and final structure of CRISPR-Cas12a along the traversal of the target strand toward the RuvC active site are available in figshare with the identifier 10.6084/m9.figshare.24978669. More details can be provided by the corresponding author upon request.

Appendix C

Supporting Information for Chapter IV

C.1 Supplementary Tables

Table C.1: **Key Resources Table**

Reagent or Resource	Source	Identifier
Bacterial and virus strains		
BL21 (DE3) One Shot Chemically Competent *E. coli*	Invitrogen	C600003
Max Efficiency DH5α-T1^R Competent Cells	Invitrogen	12034013

Continued on next page

continued from previous page

Reagent or Resource	Source	Identifier
Chemicals, peptides, and recombinant proteins		
Acidaminococcus sp. (As)Cas12a	This paper	N/A
AsCas12a Loop Deletion (Δ1049-1058)	This paper	N/A
AsCas12a ΔREC2 (Δ326-521+Gly+Ala)	This paper	N/A
AsCas12a F999A	This paper	N/A
AsCas12a R1003A	This paper	N/A
Deposited data WT Cas12a 5-bp R-loop	This paper	EMD-40441; PDB 8SFH
WT Cas12a 8-bp R-loop	This paper	EMD-40442; PDB 8SFI
WT Cas12a 10-bp R-loop	This paper	EMD-40443; PDB 8SFJ
WT Cas12a 15-bp R-loop	This paper	EMD-40444; PDB 8SFL
WT Cas12a 16-bp R-loop	This paper	EMD-40445; PDB 8SFN

Continued on next page

continued from previous page

Reagent or Resource	Source	Identifier
WT Cas12a 20-bp R-loop	This paper	EMD-40446; PDB 8SFO
WT Cas12a with TS in the active site	This paper	EMD-40447; PDB 8SFP
WT Cas12a with exposed active site	This paper	EMD-40448; PDB 8SFQ
WT Cas12a in extended conformation	This paper	EMD-40449; PDB 8SFR
Oligonucleotides		
DNA Oligos for cryo-EM, biochemistry listed in supplemental info	This paper, Tables C.3 and C.4	N/A
Cas12a minimum pre-crRNA: UUUUUAAUUUCUACUCUUGUAGAUGUGAUAAGUGGAAUGCCAUGUGGA	Strohkendl et al.[48]	N/A
Recombinant DNA		

Continued on next page

continued from previous page

Reagent or Resource	Source	Identifier
pHIS-MBP-AsCas12a	This paper	N/A
pHIS-MBP-AsCas12a-LD	This paper	N/A
pHIS-MBP-AsCas12a-ΔREC2	This paper	N/A
pHIS-MBP-AsCas12a-F999A	This paper	N/A
pHIS-MBP-AsCas12a-R1003A	This paper	N/A
pBAD-AsCas12a	This paper	N/A
pBAD-AsCas12a-LD	This paper	N/A
pBAD-AsCas12a-K1054A	This paper	N/A
pACYC-sfGFP-TargetD	This paper	N/A
pACYC-sfGFP-TargetD-crRNA	This paper	N/A
Software and Algorithms		
SerialEM	Mastronarde[181]	https://bio3d.colorado.edu/SerialEM/

Continued on next page

continued from previous page

Reagent or Resource	Source	Identifier
cryoSPARC live	Punjani et al.[302]	https://cryosparc.com/live
cryoSPARC	Punjani et al.[183]	https://cryosparc.com/
COOT	Emsley et al.[186]	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
ISOLDE	Croll et al.[185]	https://isolde.cimr.cam.ac.uk/
PHENIX	Adams et al.[187]	https://phenix-online.org/
ChimeraX	Petterson et al.[184]	http://plato.cgl.ucsf.edu/chimerax/download.html
PRISM	GraphPad	https://www.graphpad.com/

Table C.2: **Sequences of RNA (Synthego) and DNA (IDT) used for sample preparation and biochemistry reactions.** Related to Figures 4.1-4.4. Oligos labeled for cleavage and binding detection had a 5'-FAM label added.

Oligo	Sequence (5' to 3')
D_min.crRNA	UUUUUAAUUUCUACUCUUGUAGAU**GUGAUAAGU** **GGAAUGCCAUG**UGGA
TargetD_NTS	CGCTCTTCCGATCTTTTA**GTGATAAGTGGAATGCC** **ATG**TGGAGTAGCTACTGTGCT
TargetD_TS	AGCACAGTAGCTACTCCA**CATGGCATTCCACTTAT**

Continued on next page

continued from previous page

Oligo	**Sequence (5' to 3')**
	CACTAAAAGATCGGAAGAGCG
8comp_NTS	CGCTCTTCCGATCTTTTAGTGATAAGACCTTACGGT ACTGGAGTAGCTACTGTGCT
8comp_TS	AGCACAGTAGCTACTCCAGTACCGTAAGGTCTTATC ACTAAAAGATCGGAAGAGCG
12comp_NTS	CGCTCTTCCGATCTTTTAGTGATAAGTGGATACGG TACTGGAGTAGCTACTGTGCT
12comp_TS	AGCACAGTAGCTACTCCAGTACCGTATCCACTTAT CACTAAAAGATCGGAAGAGCG
16comp_NTS	CGCTCTTCCGATCTTTTAGTGATAAGTGGAATGCGT ACTGGAGTAGCTACTGTGCT
16comp_TS	AGCACAGTAGCTACTCCAGTACGCATTCCACTTATC ACTAAAAGATCGGAAGAGCG
D_TSps	AGCACAGTAGCT*A*C*T*C*C*ACATGGCATTCCAC TTATCACTAAAAGATCGGAAGAGCG
A4T_NTS	CGCTCTTCCGATCTTTTAGTGtTAAGTGGAATGCCA TGTGGAGTAGCTACTGTGCT
A4T_TS	AGCACAGTAGCTACTCCACATGGCATTCCACTTAAC ACTAAAAGATCGGAAGAGCG
ssD_20	CATGGCATTCCACTTATCAC

Continued on next page

continued from previous page

Oligo	**Sequence (5' to 3')**
ssD_16	GCATTCCACTTATCAC
trans substrate1	TTTATT
trans substrate2	CCCCCCCCCC

Table C.3: **List of observed cleavage rates measured at 37 °C.** Related to Figure 4.2 and 4.3. *Cis*-cleavage rates are listed as averages of triplicates ± standard error of the mean (Exceptions: F999A, 16-bp TS and all R1003A reactions were performed in duplicate). *Trans* second order cleavage rates are listed as averages of at least duplicates ± standard error of the mean. *Trans* cleavage reactions with no observed cleavage were assigned an upper limit to cleavage of 350 M^{-1} min^{-1}. Competition assays were measured in duplicate and rates and final amplitudes are listed as averages ± standard error of the mean. '—' marks rates that were not measured.

Cas12a, R-loop	**NTS** (min^{-1})	**TS** (min^{-1})	***trans*** (M^{-1} min^{-1})
WT, 20bp	6.25 ± 0.73	0.782 ± 0.075	814000 ± 111000
WT, 16bp	1.32 ± 0.31	0.1188 ± 0.0056	38000 ± 5700
WT, A4T	0.172 ± 0.015	—	—
LD, 20bp	5.03 ± 0.44	—	—
LD, A4T	0.080 ± 0.016	—	—
ΔRec2, 20bp	0.0963 ± 0.0081	0.00103 ± 0.00053	—
ΔRec2, 16bp	0.0158 ± 0.0089	0.0037 ± 0.0024	—

Continued on next page

continued from previous page

Cas12a, R-loop	**NTS** (min^{-1})	**TS** (min^{-1})	***trans*** (M^{-1} min^{-1})
F999A, 20bp	4.57 ± 0.14	0.0857 ± 0.0073	350
F999A, 16bp	0.0521 ± 0.0070	0.0009 ± 0.0002	350
R1003A, 20bp	5.52 ± 0.81	0.547 ± 0.053	—
R1003A, 16bp	0.097 ± 0.019	0.0181 ± 0.0045	—

Table C.4: **Competitor Cleavage Rates**

Competitor, incubation	**Rate** (min^{-1})	**Fraction cleaved**
8bp, 1 hr	7.6 ± 1.1	0.97 ± 0
12bp, 1 hr	6.61 ± 0.06	0.97 ± 0.01
16bp, 30 sec	2.2 ± 0.3	0.93 ± 0.03
16bp, 2 min	0.52 ± 0.02	0.65 ± 0.03
20bp, 30 sec	0.03 ± 0.01	0.1 ± 0.01

C.2 Supplementary Figures

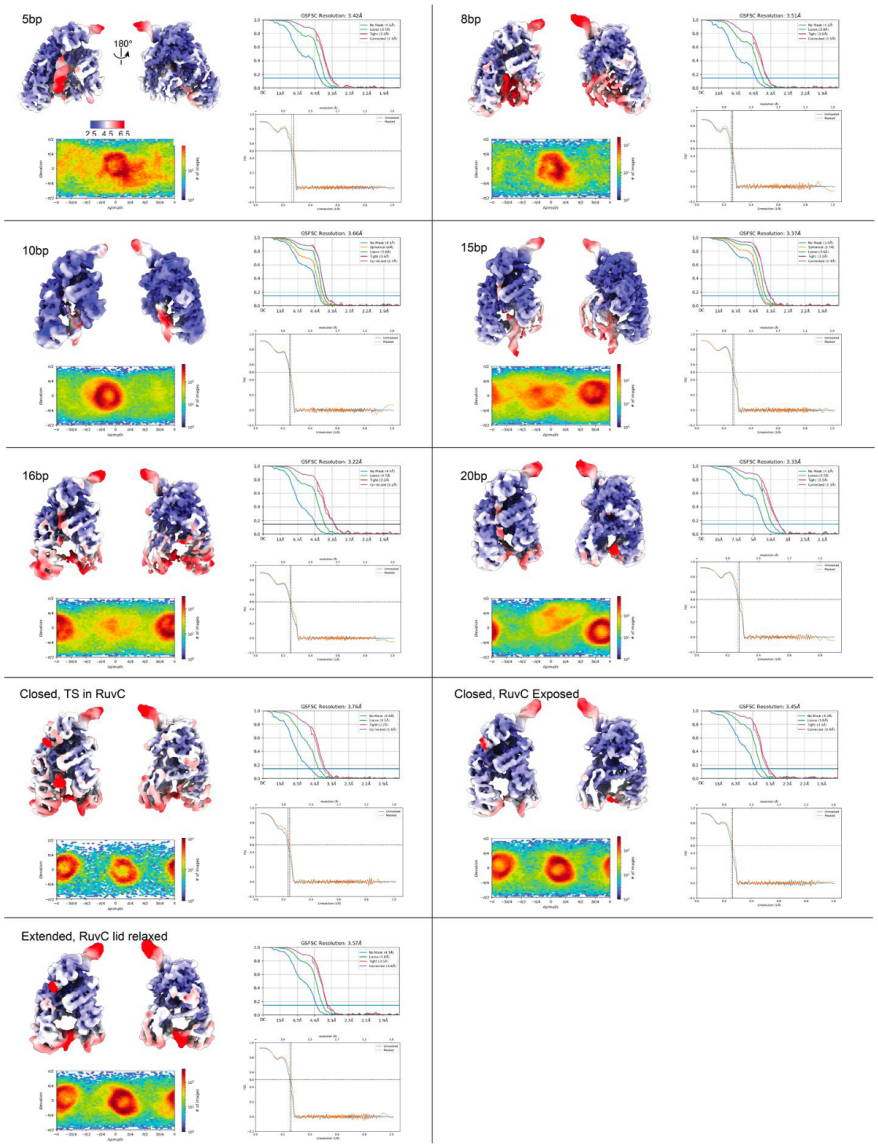


Figure C.1: **Local resolution of intermediate R-loop maps.** Related to Figures 4.1 and 4.4 and Table 4.1. For each panel, in clockwise order: unsharpened maps colored by local resolution. (as portrayed in Figures 4.1 and 4.4), reconstruction gold-standard FSC curves, resolution estimated at FSC = 0.143, reconstruction Euler diagrams showing particle orientation distributions, and map-to-model FSC curves produced in Phenix. Map resolutions and particle distribution plots were produced in cryoSPARC.

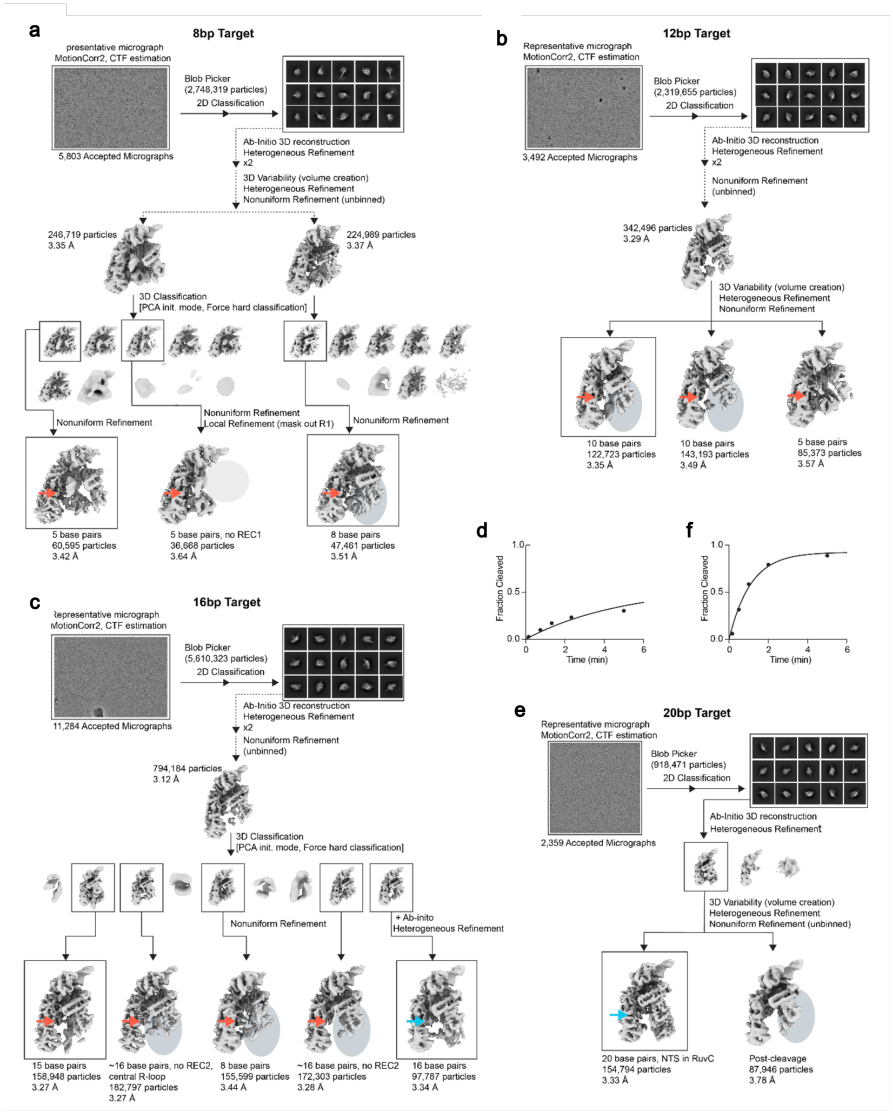


Figure C.2: **Data processing pipelines for Cas12a R-loop datasets.** Related to Figure 4.1. **(a)** 8-bp dataset. **(b)** 12-bp dataset. **(c)** 16-bp dataset. **(d)** Example time course of NTS cleavage of the 16-bp target at ambient temperatures used for preparing the cryo-EM sample. **(e)** 20-bp dataset. **(f)** Example time course of NTS cleavage of the 20-bp target at ambient temperatures used for preparing the cryo-EM sample. At one minute, nearly 50% cleavage is observed and suggests the second structural class represents a post-cleavage state. Final boxed reconstructions are reported in Figures 4.1 and 4.4. Grey circles highlight flexible and unresolved Rec domains. Arrows pointing to the RuvC active site mark whether it is occluded (red) or bound to NTS (light blue). Representative structures during processing are shown after reaching a consensus reconstruction.

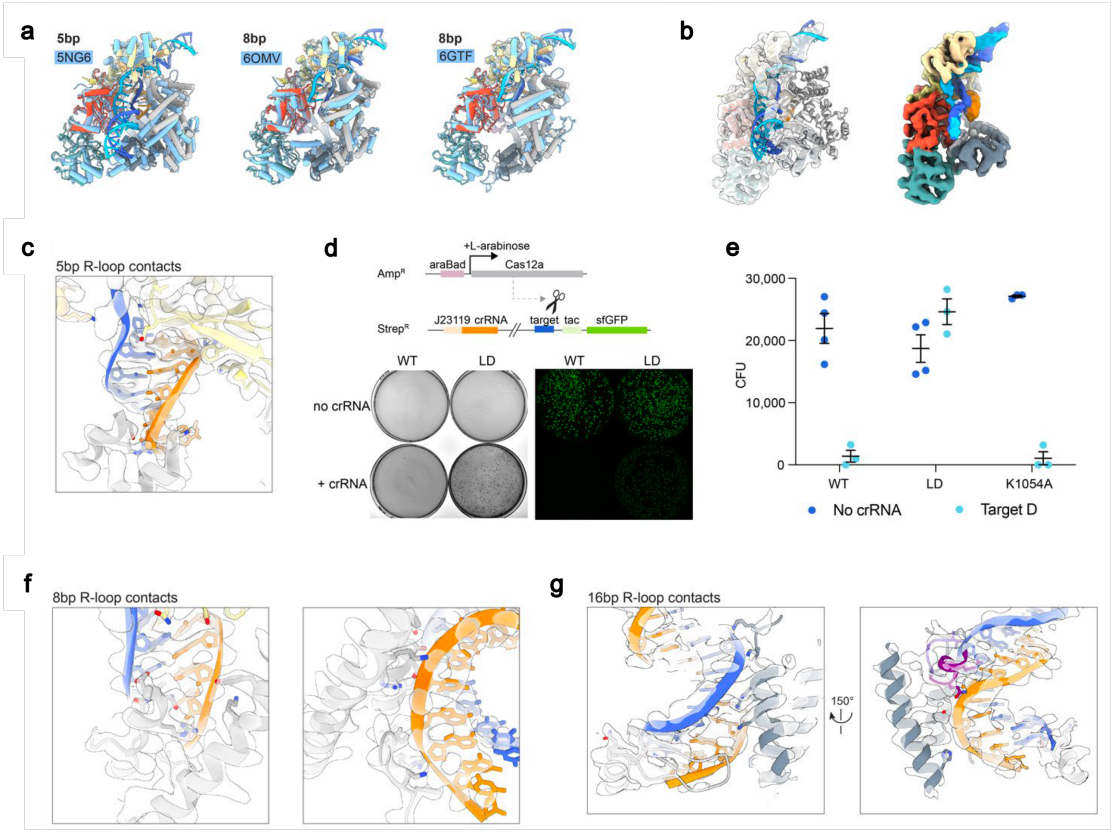


Figure C.3: **Cas12a early and middle R-loops.** Related to Figure 4.2. **(a)** Comparison of cryo-EM structures to previously published binary and 8-bp structures from Fn and Lb Cas12a structures. **(b)** Map- model overlay of the final 5-bp R-loop structure. **(c)** Left: The 5-bp reconstruction lacking a resolved REC1 domain is overlayed with the complete 5-bp structural model. Right: The 5-bp reconstruction lacking REC1 is colored according to Figure 4.1. **(d)** Experimental design for plasmid interference assay in *E. coli*. BL21 DE3 cells were co-transformed with plasmids expressing Cas12a effector and crRNA targeting a sequence within the plasmid enabling antibiotic selection. Before plating, expression of Cas12a was induced with the addition of 2% L-arabinose. Bottom: example images (bright light and GFP-fluorescence) of plates selecting for *E. coli* cells with ampicillin- and streptomycin-resistance. **(e)** Quantification of CFUs per plate per condition. Error bars represent standard error of the mean (SEM). **(f)** Map-model overlays of the final 8-bp R-loop structure. **(g)** Map-model overlays of the 16-bp R-loop contacts. Sharpened maps are used for zoomed-in overlays.

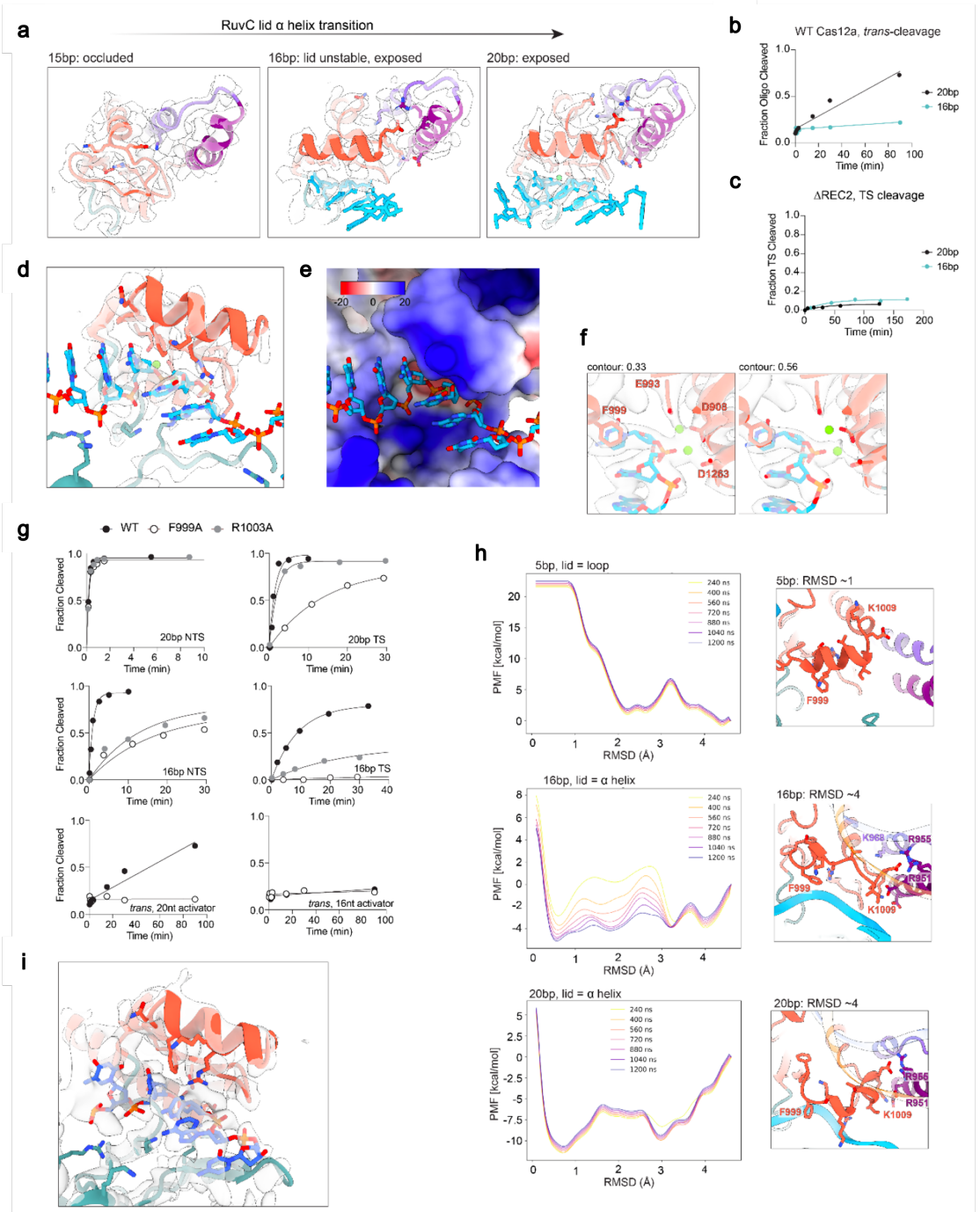

a
RuvC lid α helix transition
15bp: occluded
16bp: lid unstable, exposed
20bp: exposed
b
WT Cas12a, trans-cleavage
Fraction Oligo Cleaved
Time (min)
20bp
16bp
c
ΔREC2, TS cleavage
Fraction TS Cleaved
Time (min)
d
e
-20
0
20
f
contour: 0.33
contour: 0.56
E993
F999
D908
D1263
g
WT
F999A
R1003A
Fraction Cleaved
Time (min)
20bp NTS
20bp TS
16bp NTS
16bp TS
trans, 20nt activator
trans, 16nt activator
h
5bp, lid = loop
PMF [kcal/mol]
RMSD (Å)
240 ns
400 ns
560 ns
720 ns
880 ns
1040 ns
1200 ns
5bp: RMSD ~1
K1009
F999
16bp, lid = α helix
16bp: RMSD ~4
K968
R955
R951
20bp, lid = α helix
20bp: RMSD ~4
i

Figure C.4: **Cas12a RuvC activation.** Related to Figures 4.3 and 4.4. **(a)** Map-model overlays of the 15-bp, 16-bp, and 20-bp structures highlighting the RuvC lid and BH structural changes. Sharpened maps are used for overlays. **(b)** Example time course of *trans*-cleavage by WT Cas12a. **(c)** Example time course of TS cleavage by ΔREC2 Cas12a mutant. **(d)** Map-model overlay of the 20-bp structure highlighting the NTS interacting with the RuvC lid. **(e)** Surface electrostatic potential representation of panel **(d)**. **(f)** Map-model overlays of the 20-bp RuvC active site shown at two different contour thresholds. Continuous density is consistent with a two Mg^{2+} ion mechanism. **(g)** Example cleavage time courses for WT, F999A, and R1003A Cas12a enzymes. Each plot is labeled according to the R-loop length and cleavage substrate being measured. **(h)** Convergence of the potential of mean force (PMF) along the reaction coordinate (RMSD), capturing the transition of the RuvC lid from α-helix to unstructured loop during different stages of R-loop formation (5-bp, 16-bp and 20-bp). To the right of each plot is a zoomed-in view of the distorted local minima. The R-loop TS and crRNA backbones are shown in transparent cartoon for clarity. The RuvC lid faces the R-loop minor groove. As the 5-bp loop transitions to the α-helix, BH contacts are lost but K1000 maintains its contact with catalytic E993. As the 16-bp α-helix transitions to an unstructured loop, the RuvC lid F999 and R1003 lose contact with the NTS, K1000 contacts E993, and BH contacts are maintained. As the 20-bp α-helix transitions to an unstructured loop, the RuvC lid remains contacting the NTS and BH and K1000 does not form a contact to sequester E993. **(i)** Map-model overlay of the TSps structure highlighting the TS interacting with the RuvC lid.

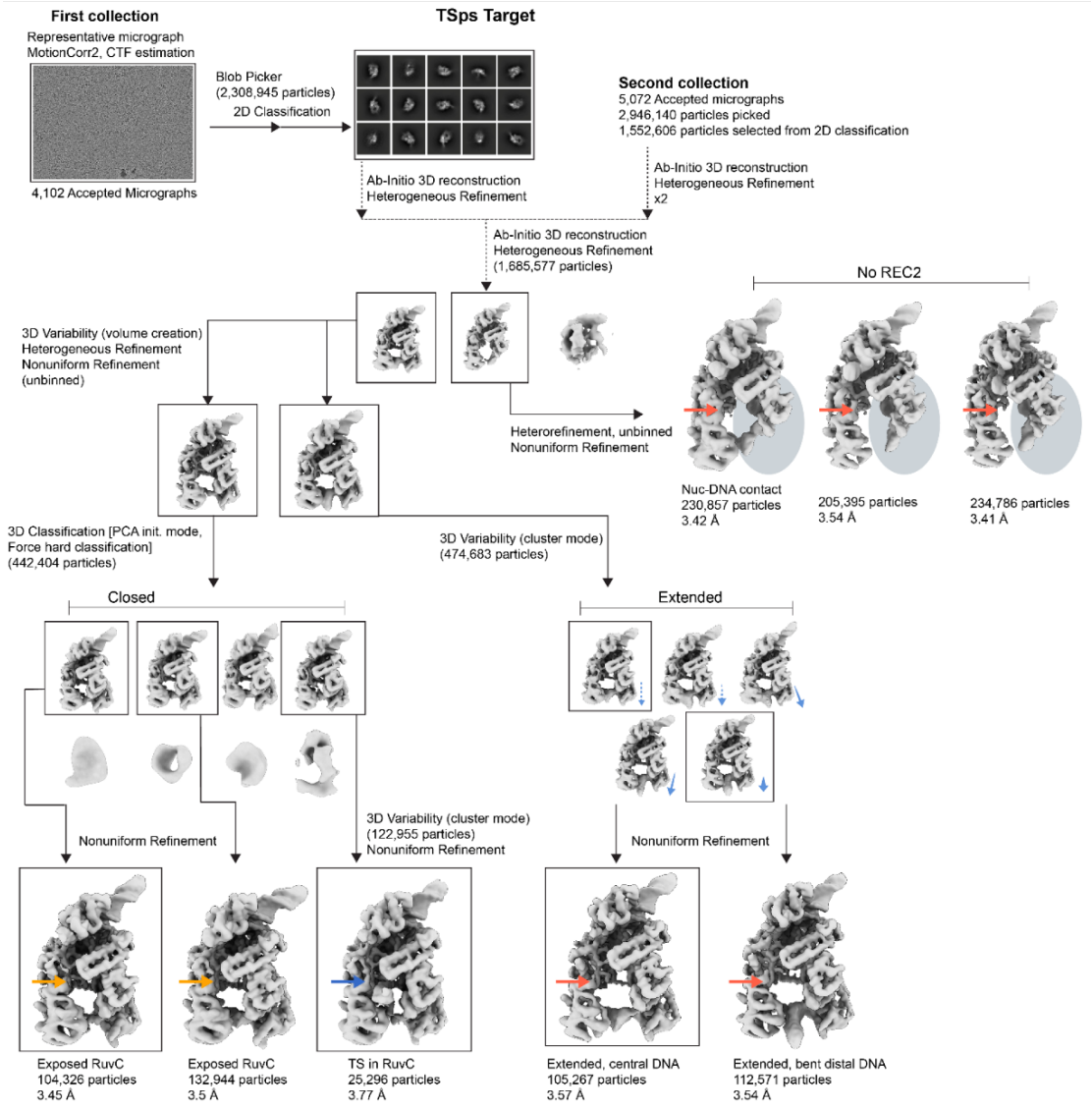


Figure C.5: **Data processing pipeline for Cas12a post-NTS cleavage dataset.** Related to Figure 4.4. Final boxed reconstructions are reported in Figure 4.4. Arrows pointing to the RuvC active site mark whether it is occluded (red), bound to TS (dark blue), or exposed (yellow). Representative structures during processing are shown after reaching a consensus reconstruction. Grey circles highlight flexible and unresolved Rec domains. Reconstructions within the "extended" group have blue arrows highlighting the variability in the distal DNA.

C.3 Data and code availability

- Structures of the 5-bp, 8-bp, 10-bp, 15-bp, 16-bp, 20-bp R-loop intermediates have been deposited in the EMDB with accession codes EMD-40441, EMD-40442, EMD-40443, EMD-40444, EMD-40445, EMD-40446, respectively. Associated atomic coordinates are deposited with accession codes PDB: 8SFH, 8SFI, 8SFJ, 8SFL, 8SFN, 8SFO, respectively. Structures from the TSps dataset have been deposited in the EMDB EMD-40447, EMD-40448, EMD-40449. Associated atomic coordinates are deposited with accession codes PDB: 8SFP, 8SFQ, 8SFR.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

C.4 Bacterial Strains and Growth Conditions

E. coli DH5α T1R competent cells (Invitrogen, 12297016) were used for cloning, following manufacturer's protocol. Recombinant Cas12a proteins were expressed in chemically-competent *E. coli* BL21(DE3) cells (Invitrogen, C600003) or NiCo21(DE3) cells (NEB, C2529H) in LB media at 18 °C. Plasmid interference assays were performed in BL21(DE3) cells (Invitrogen) grown in SOC media at 37 °C before plating on LB agar plates supplemented with antibiotics and chemicals for expression induction.

C.5 Limitations of the Study

Because our prior kinetic characterization of Cas12a identified a late transition state to fast R-loop formation, we reasoned intermediate R-loop states would be too short-lived to capture using a matched target DNA. Thus, to favor observation of various intermediate states that sample the conformational landscape of Cas12a DNA targeting, we used a series of DNA substrates with varying amounts of complementarity before reaching the transition state for binding. As a result, some structural datasets represent binding equilibria that change the relative populations of R-loop intermediates compared with those expected for the on-target substrate. Lastly, structures capturing Cas12a PAM recognition and pre-seed R-loops are still needed for a complete understanding of DNA targeting.

Appendix D

Supporting Information for Chapter V

D.1 Supplementary Tables

Table D.1: **PAGE analysis of LbCas12a *trans*-cleavage activity.** The fraction of DNA reporters cleaved by the LbCas12a/crRNA complex activated by dsDNA vs time and resolved by Polyacrylamide gel electrophoresis (PAGE) are reported as cleavage %. Samples were prepared by adding a 10X concentrated solution of RNP complex (Cas12a 200 nM/crRNA 200 nM/dsDNA activator 10 nM) to a solution containing the DNA reporter (final concentration 250 nM) at 37°C. The intensity is measured by stopping the *trans*-cleavage reaction at different times (0, 15, 30, 45, 60 and 120 min), by heating the solutions for 10 min at 75 °C to inactivate Cas12a enzyme.

	Linear ss-DNA	**Stem-loop #5T**	**Stem-loop #10T**	**Stem-loop #30T**
Time	**Cleavage % (PAGE assay)**	**Cleavage % (PAGE assay)**	**Cleavage % (PAGE assay)**	**Cleavage % (PAGE assay)**
15 min	18	22	33	42
30 min	34	43	71	55

Continued on next page

continued from previous page

	Linear ss-DNA	**Stem-loop #5T**	**Stem-loop #10T**	**Stem-loop #30T**
Time	**Cleavage % (PAGE assay)**	**Cleavage % (PAGE assay)**	**Cleavage % (PAGE assay)**	**Cleavage % (PAGE assay)**
45 min	40	61	82	79
1 h	65	70	87	83
2 h	100	100	100	100

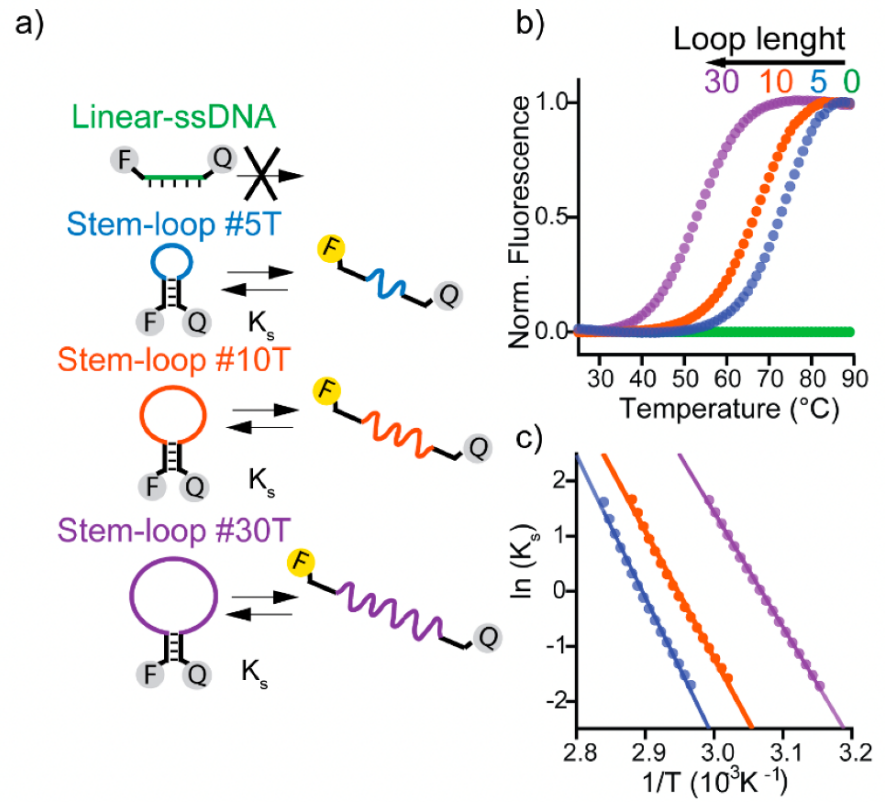
Table D.2: **Measured K_m, k_{cat}, and k_{cat}/K_m.** The values are reported as mean ± SD, where n = 3.

	K_m **(nM)**	k_{cat} **(s^{-1})**	k_{cat}/K_m **($s^{-1}M^{-1}$)**
Linear ss-DNA	250 ± 50	0.30 ± 0.02	(1.2 ± 0.2) x 10^6
Stem-loop #5T	42 ± 5	0.24 ± 0.02	(6.1 ± 0.9) x 10^6
Stem-loop #10T	34 ± 2	0.23 ± 0.01	(6.7 ± 0.7) x 10^6
Stem-loop #30T	69 ± 5	0.28 ± 0.02	(4.1 ± 0.6) x 10^6

Table D.3: **Comparison of enzyme kinetic parameters.** The most relevant published data reported in literature for LbCas12a and dsDNA activator is also provided.

Reporter	K_m **(nM)**	k_{cat} **(s^{-1})**	k_{cat}/K_m **($s^{-1}M^{-1}$)**	**Reference**
DNaseAlert	1010	17	1.71 x 10^7	Chen et al., 2021, Science, 371, No. eabh0317
DNaseAlert	26	1.30	5.1 x 10^7	Nalefski et al., 2021, iScience, 24, 102996
/56-FAM/TTA TTATT/BHQ-1	213 274	0.09 0.07	4.2 x 10^5 2.5 x 10^5	Ramachandran et al., 2021, Anal. Chem., 93, 7456–7464

D.2 Supplementary Figures



	T_m(°C)	ΔG(Kcalmol^{-1})	K_s
Stem-loop #5T	72.4 ± 0.3	4.5 ± 0.2	(7 ± 2) × 10^{-4}
Stem-loop #10T	67 ± 1	3.9 ± 0.1	(1.7 ± 2.2) × 10^{-3}
Stem-loop #30T	53.3 ± 0.5	2.19 ± 0.015	(2.9 ± 0.2) × 10^{-2}

Figure D.1: **Thermodynamic characterization of the FRET-based hairpin DNA reporters.** **(a)** Schematic representation of the structure-switching mechanism occurring between the closed and opened states of the hairpin structures. The Linear-ssDNA reporter does not show any conformational equilibrium. **(b)** Melting curve and **(c)** van't Hoff analysis of the melting assays carried out to provide the quantitative information reported in table about the switching equilibrium constants (K_s) associated to the specific hairpin-forming DNA reporters.

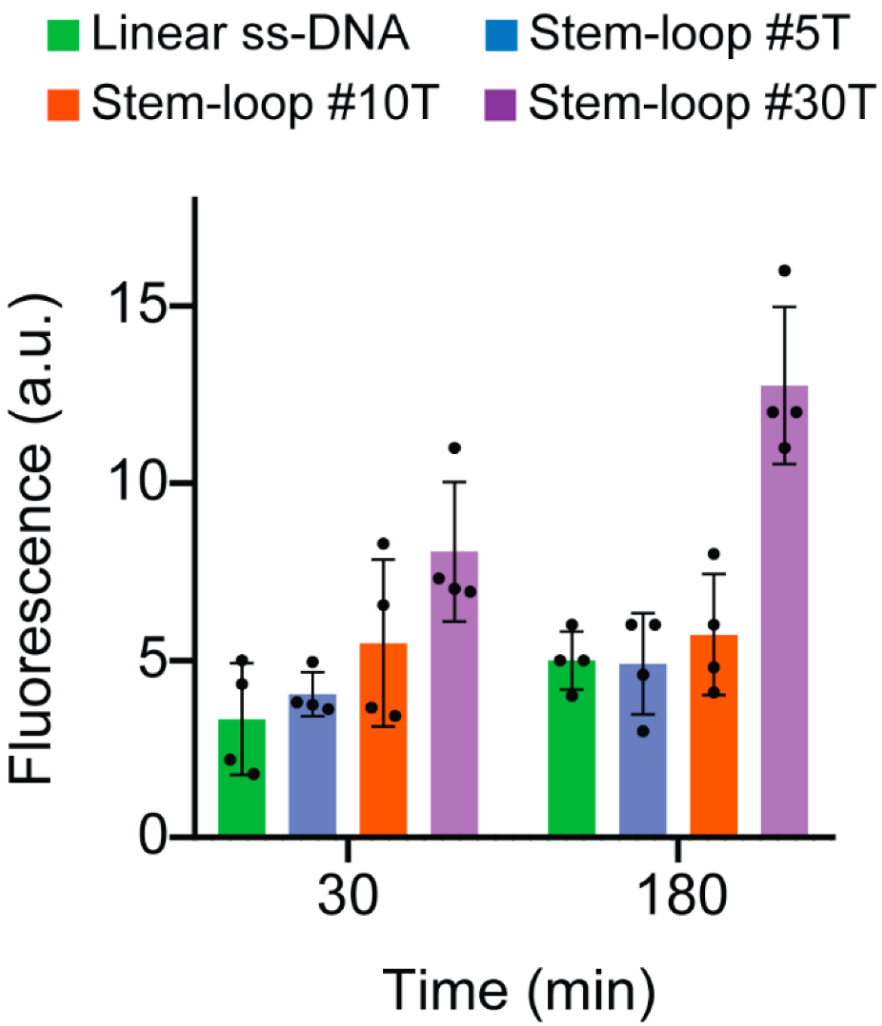


Figure D.2: **Background fluorescence of the DNA reporters.** Fluorescence background of the DNA reporters (100 nM) at 37°C after 30 and 180 min showing very low background change over time and higher background of fluorescence for Stem-loop #30T according to the fraction of this reporter in the unfolded state at temperature of the assays (T = 37 °C).

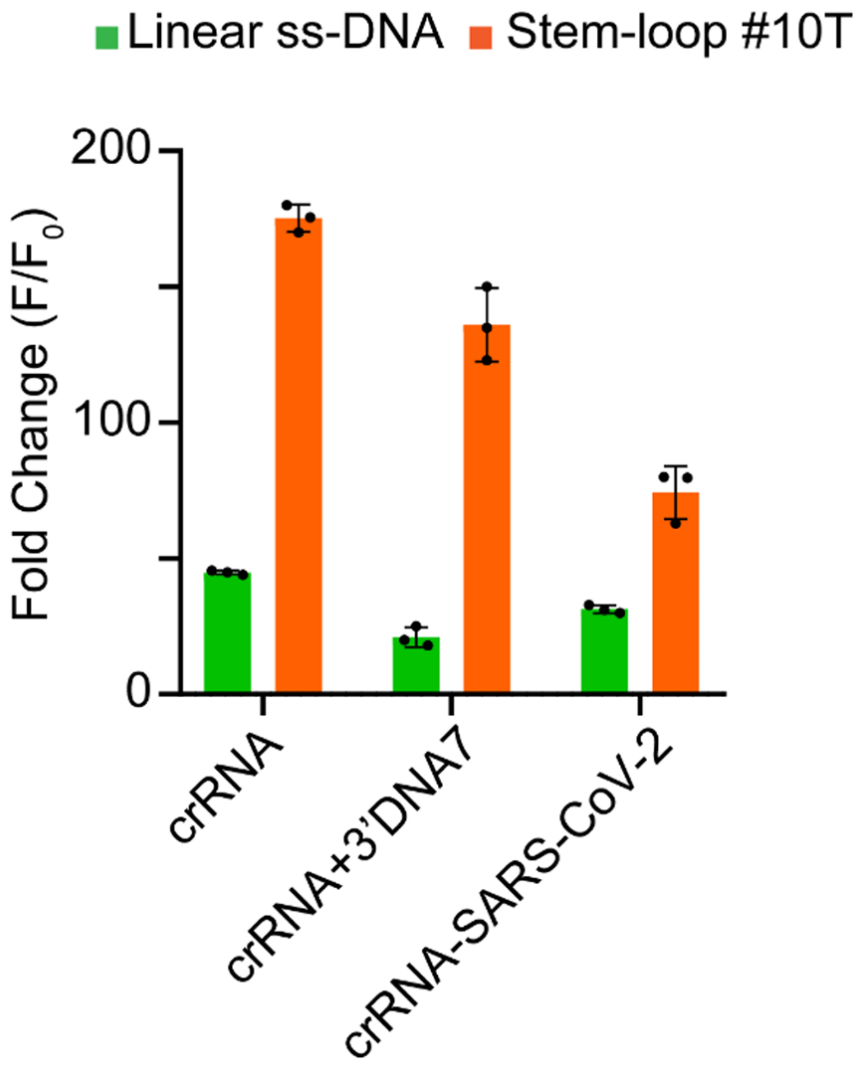


Figure D.3: ***Trans*-cleavage activity was tested by using different crRNA and dsDNA activators (see Oligonucleotides in Materials and Methods).** Fluorescence is plotted as F/F_0 (F is the fluorescence signal subtracted of the background F_0). F_0 is the fluorescence background in the absence of the dsDNA activator collected at time t = 60 min. The bar graph shows the mean and SD as error bars (n = 3) at the assay endpoint (t = 60 min). All the experiments were carried out by adding 4.5 μl of 10x concentrated solution of RNP complex to a solution (final volume 45 μl) containing the FRET-based DNA reporter (final concentrations: Cas12a/crRNA/dsDNA activator 20/20/1 nM, [reporter] = 100 nM).

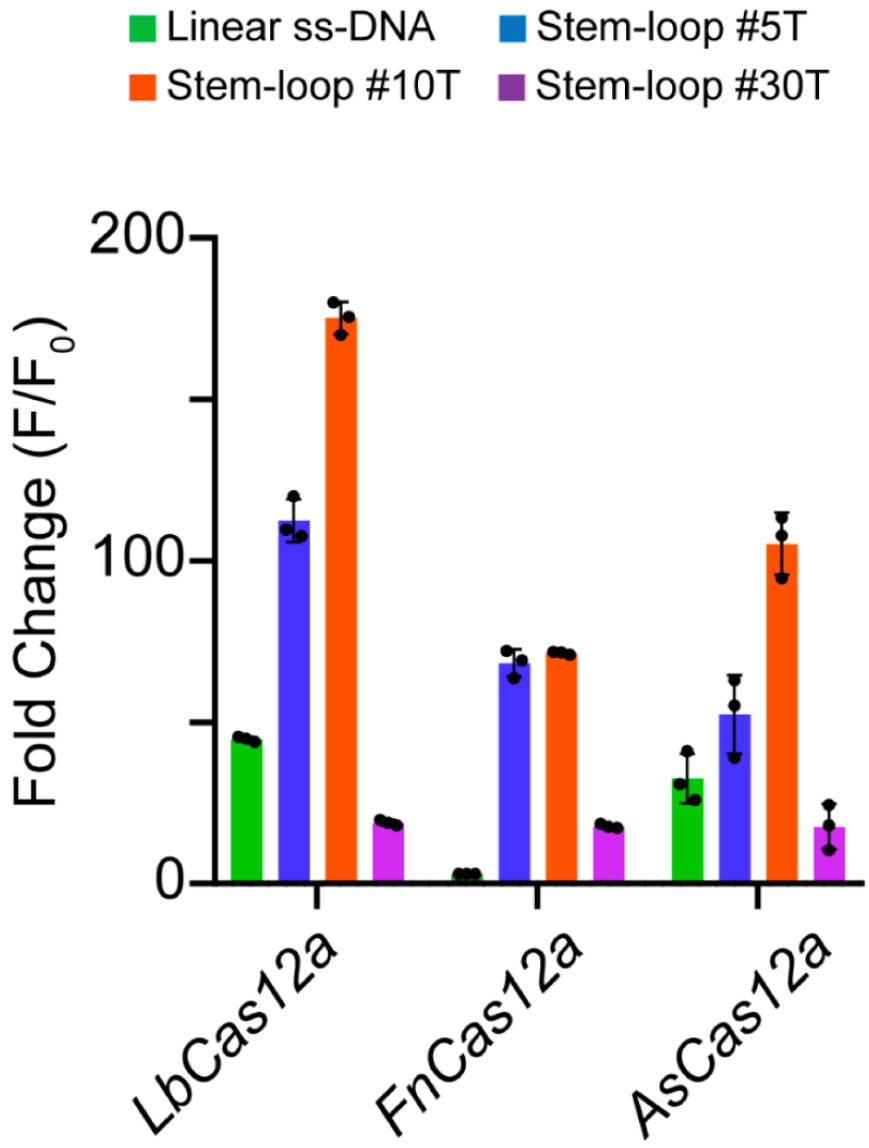


Figure D.4: ***Trans*-cleavage activity was tested using different orthologs of Cas12a enzyme.** Fluorescence is plotted as F/F_0 at time t = 60 min (F_0 is fluorescence background measured in the absence of the dsDNA activator). The bar graph shows the mean and SD as error bars (n = 3) at the assay endpoint (t = 60 min). All the experiments were carried out by adding 4.5 μl of 10x concentrated solution of RNP complex to a solution (final volume 45 μl) containing the FRET-based DNA reporter (final concentrations: Cas12a/crRNA/dsDNA activator 20/20/1 nM, [reporter] = 100 nM).

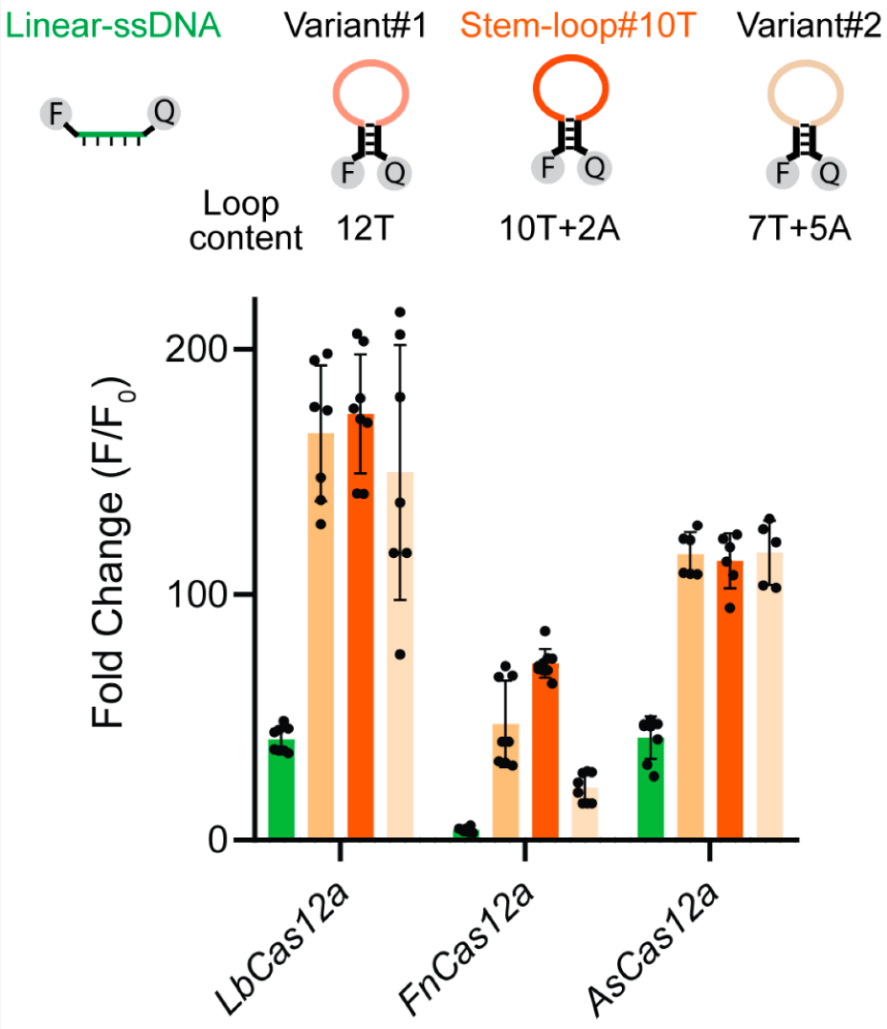


Figure D.5: ***Trans*-cleavage activity of different Cas12a orthologs using variants of Stem-loop #10T reporter.** The three hairpin reporters have a different number of thymine in the loop but they show very similar fluorescence fold change. Fluorescence is plotted as F/F_0 at time t = 60 min (F_0 is fluorescence background measured in the absence of the dsDNA activator). The bar graph shows the mean and SD as error bars (n = 7) at the assay endpoint (t = 60 min). All the experiments were carried out by adding 4.5 μl of 10x concentrated solution of RNP complex to a solution (final volume 45 μl) containing the FRET-based DNA reporter (final concentrations: Cas12a/crRNA/dsDNA activator 20/20/1 nM, [reporter] = 100 nM).

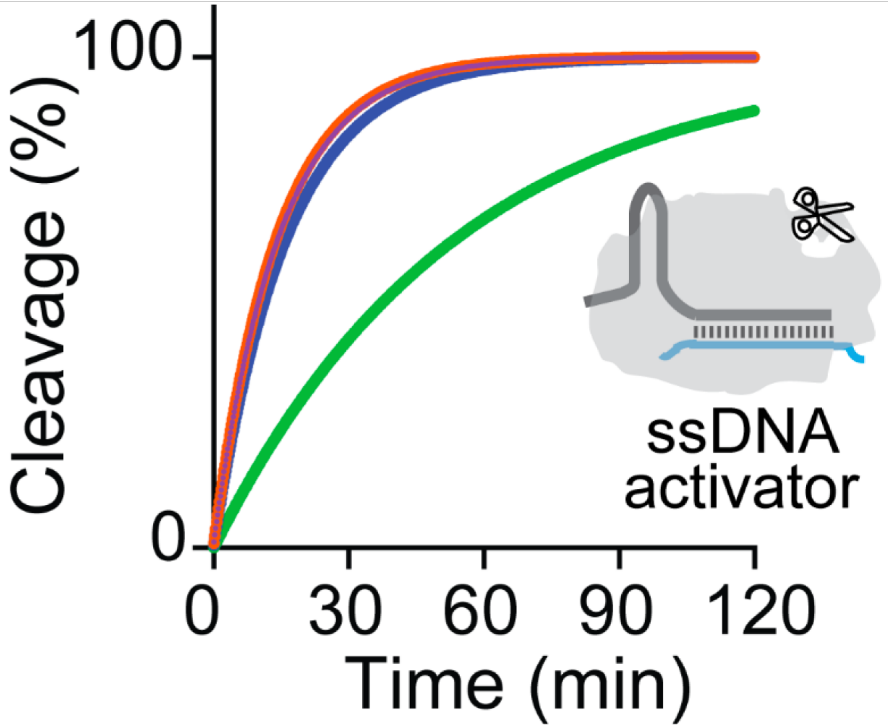


Figure D.6: **Kinetic analysis of LbCas12a *trans*-cleavage activity.** Fraction of DNA reporters cleaved by the LbCas12a/crRNA complex activated by ssDNA activators vs time. The assays were performed at 37°C by adding 4.5 μl of Cas12a/crRNA complex (200 nM) preassembled with the ssDNA activator (10 nM) to a solution containing the DNA reporter (100 nM).

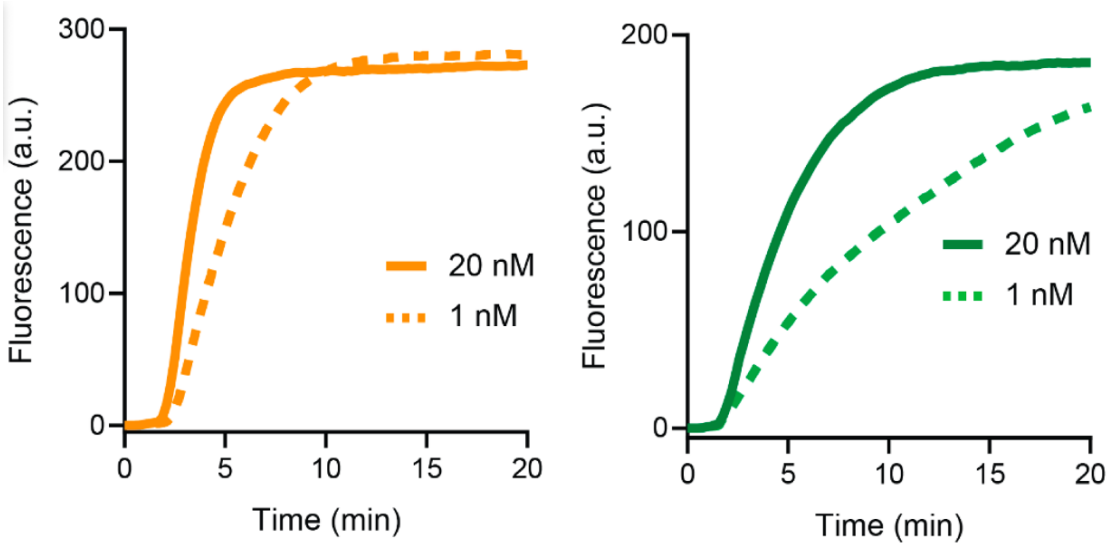


Figure D.7: **Kinetic analysis of LbCas12a *trans*-cleavage activity on Stem-loop #10T (left) and Linear ss-DNA (right) reporters.** These reporters are tested at high (20 nM) and low concentration (1 nM) of ds-DNA activator. The assays were performed at 37 °C by adding a 10X concentrated solution of Cas12a/crRNA complex (200 nM) preassembled with the dsDNA activator (10 nM) to a solution containing the DNA reporter (final concentration 100 nM).

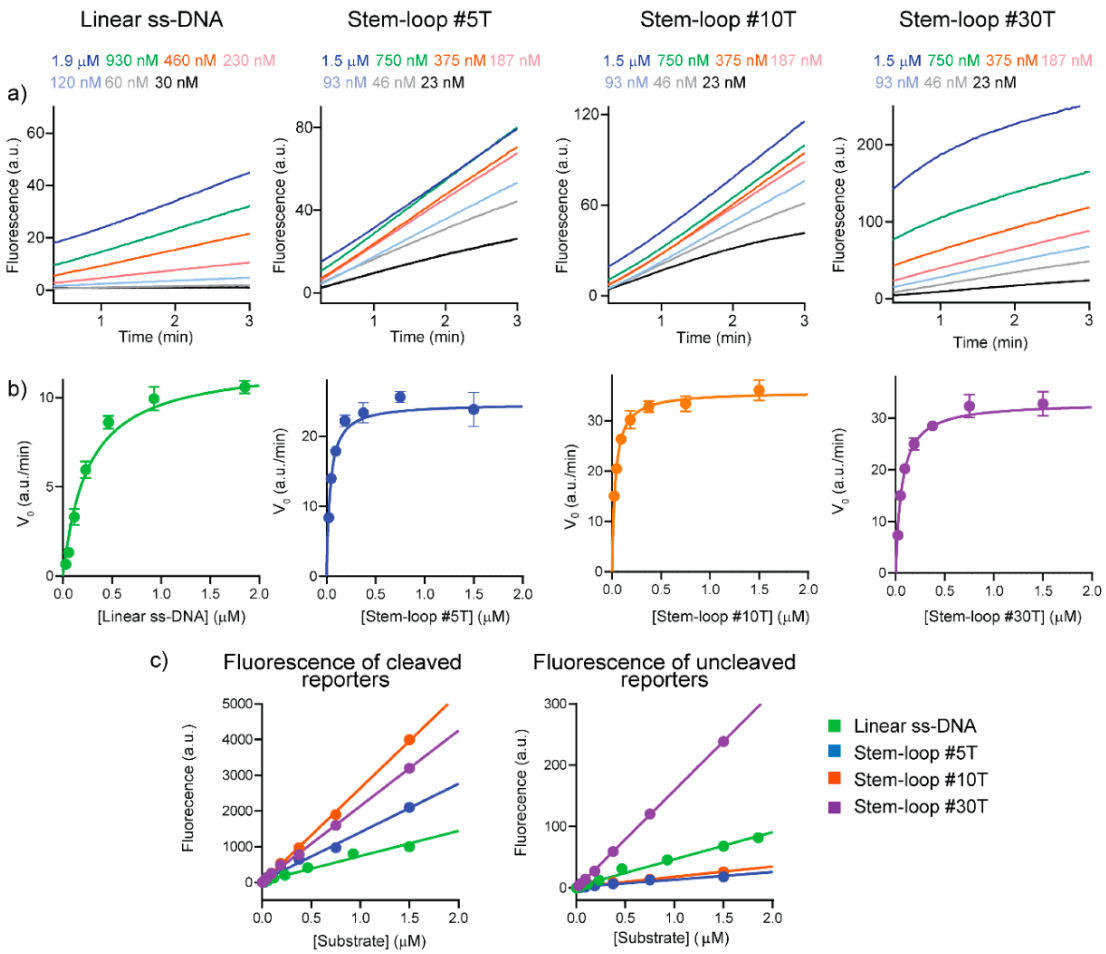


Figure D.8: **Michaelis-Menten analysis.** **(a)** Representative plots of initial velocity versus time using a fixed concentration of LbCas12a/crRNA/dsDNA activator ribonucleoprotein complex (1nM) and different reporter substrate concentrations. **(b)** Michaelis-Menten plots for the corresponding data reported in (a). **(c)** Calibration curves of cleaved and uncleaved reporters used to convert reaction velocities from a.u./min to nM/min. All the values reported represent the mean and SD as error bars (n = 3).

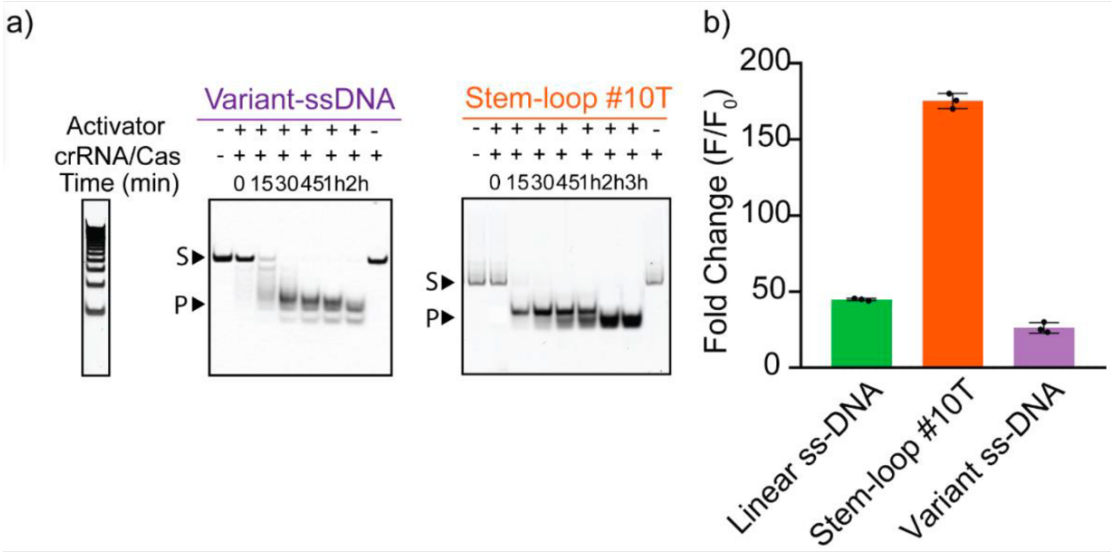


Figure D.9: ***Trans*-cleavage activity of Cas12a on Variant-ssDNA and Stem-loop #10T reporters.** **(a)** PAGE assay showing the time-dependent *trans*-cleavage activity of the activated RNP complex on FRET-labelled DNA reporters (250 nM). S represent the substrate (e.g. DNA reporters, S) and P the degraded products P. The intrinsic fluorescence background of the DNA reporters (lines 1 and 2) and signal change over time is consistent with the increased amount of degraded oligonucleotides over time (lines 3 to 7). **(b)** Fluorescence fold change after 1h of *trans*-cleavage activity indicating enhanced signal transduction using Stem-loop #10T reporter.

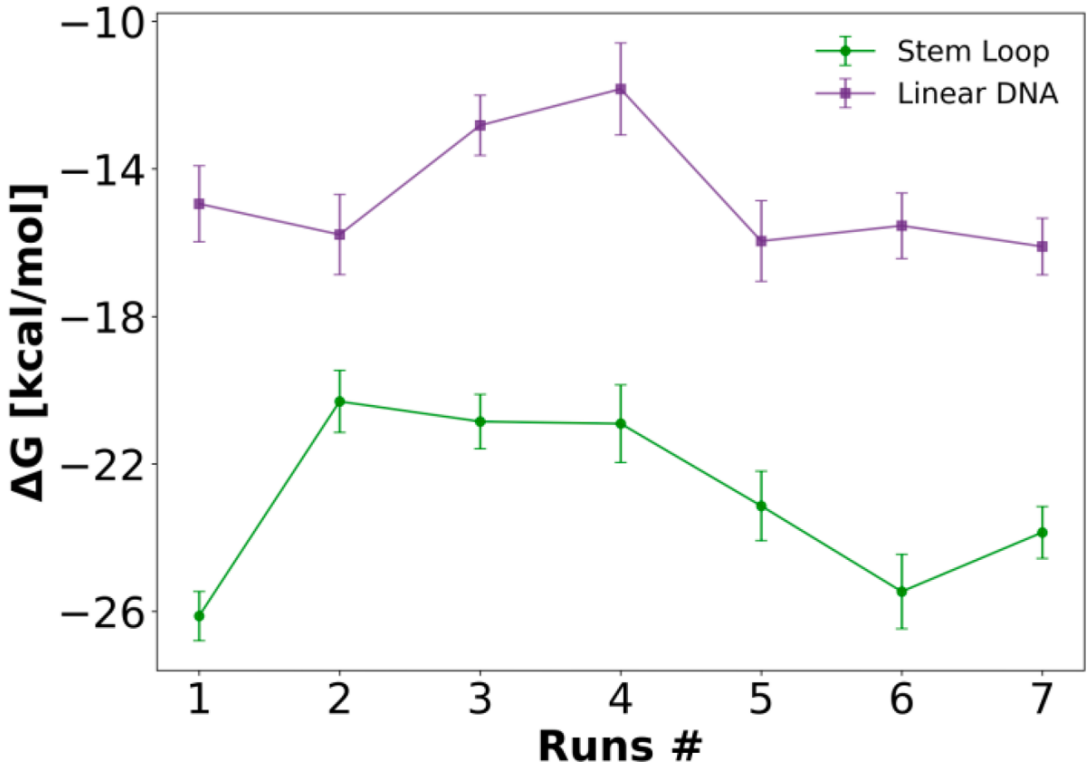


Figure D.10: **The binding free energy for the stem loop (green) and linear ssDNA(purple) binding Cas12a.** Data are shown for the individual runs of the MM-GBSA solvation method, showing stronger association of the stem loop with Cas12a than the linear ssDNA. The average free energy of the individual runs is plotted with the standard error of mean.

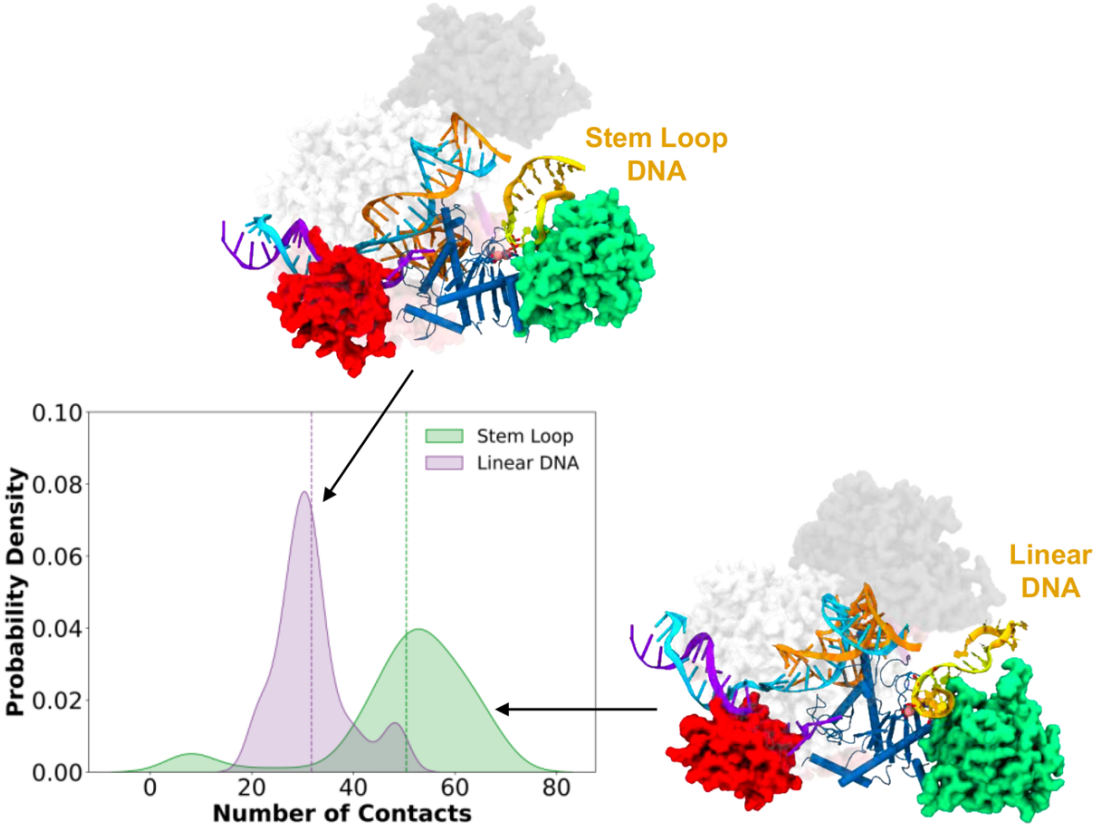


Figure D.11: **Contact analyses of the *trans*-cleavage substrates (i.e., stem loop and linear DNA) with the Nuc domain along the MD simulations.** The shift in the probability density plots (dashed lines representing the mean) shows larger extent of interactions between the stem loop DNA and the Nuc domain, compared to the linear DNA. Contacts were considered for interactions within 5 Å, to account for both strong and weak interactions.

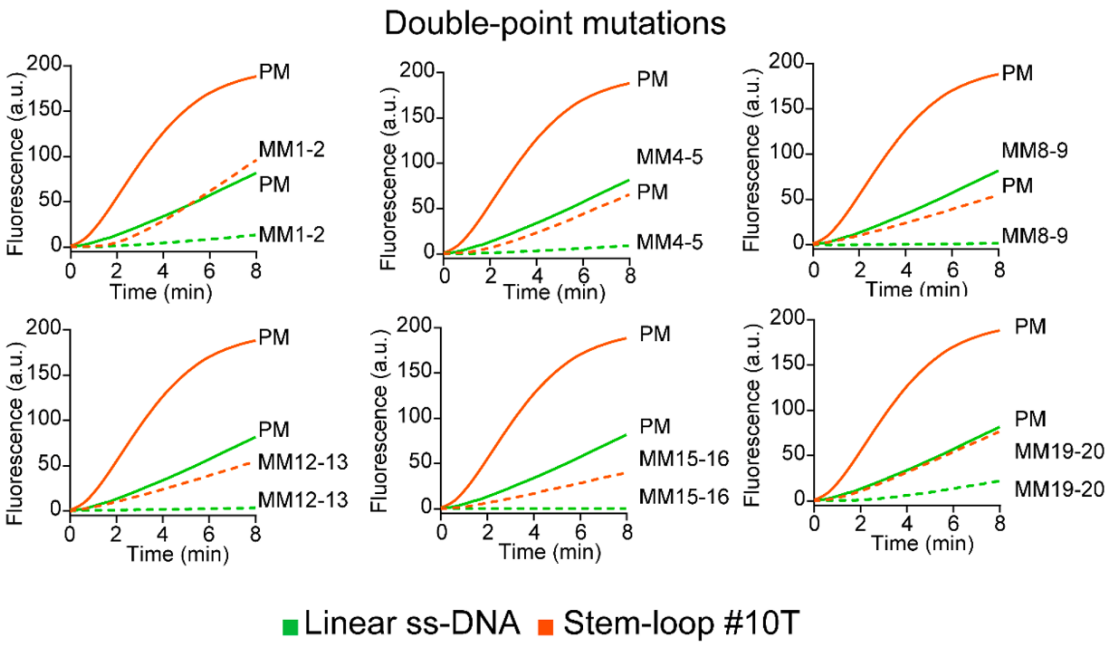


Figure D.12: ***Trans*-cleavage reaction profile over time to evaluate the specificity of LbCas12a in the presence of wilt type and mutant dsDNA target.** All the experiments were carried out by adding 4.5 μl of dsDNA (final concentration 10 nM) to 45 μl of solution containing the DNA reporter (final concentration 100 nM) and 20 nM of LbCas12a/crRNA complex.

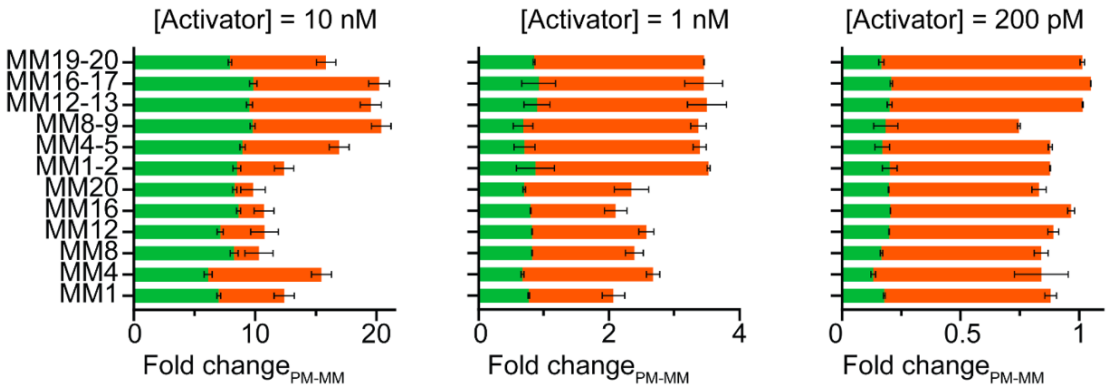


Figure D.13: **Specificity test on single-nucleotide mismatch mutations (from MM1 to MM20) and double- nucleotide mismatch mutations (from MM1-2 to MM19-20) on synthetic SARS-CoV-2 dsDNA activator at different concentrations (10 nM, 1 nM, 200 pM).** Superimposed bar graphs showing the difference between the fold change related to the perfect match (PM) and the fold change related to the mismatch (MM) target (right, t = 4 min). All the values reported show the mean ± SD, where n = 3 replicates.

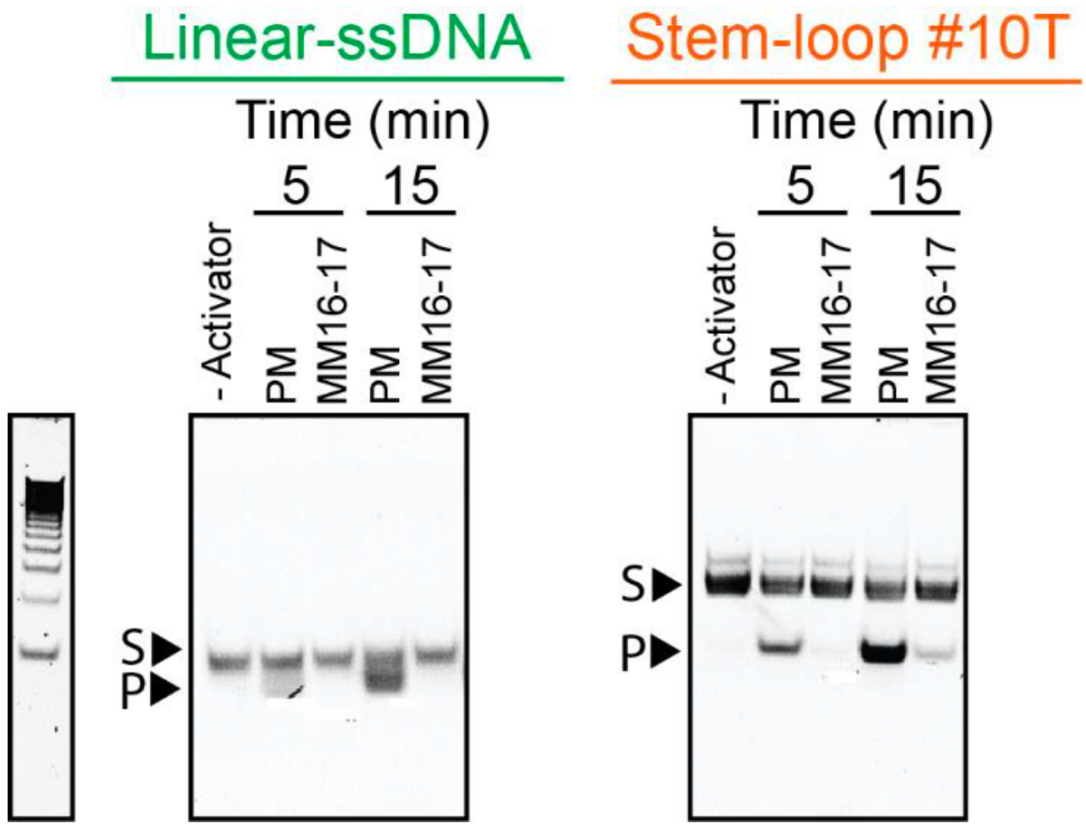


Figure D.14: **Time-dependent *trans*-cleavage activity of LbCas12a activated by native ds-DNA (PM) and double-point mutations ds-DNA (MM16-17) activators.** PAGE assay showing the time-dependent *trans*-cleavage activity of activated LbCas12a on Linear-ssDNA and Stem-loop #10T reporters (250 nM). S indicate the substrate (e.g. DNA reporters) whereas P indicate the degraded products. The intrinsic fluorescence background of the DNA reporters (lines 1) and signal changes at 5 and 15 min indicating the enhanced specificity obtained using Stem-loop #10T reporter.

Appendix E

Supporting Information for Chapter VI

E.1 Structural Models

Molecular simulations were based on the cryo-EM structure of the ABE8e complex, composed of a catalytically impaired CRISPR-Cas9 system (dCas9) from *S. pyogenes* and the TadA8e enzyme solved at 3.20 Å resolution (PDB: 6VPC).[259] Based on this cryo-EM structure, which preserves the *bona fide* configuration of the TadA enzymes bound to CRISPR-Cas9, four model systems of the ABE complexes were built: ABE0.1 (where TadA is WT), ABE7.10, ABE8e, and an additional ABE8e including the TadA8e heterodimer, in which the catalytic domain is the evolved TadA8e and the docking domain is the wt-TadA (hereafter referred as ABE8e hetero). The X-ray structure of the wt-TadA (PDB ID: 1Z3A,[261] at 2.03 Å resolution) was used to model the wt-TadA and the docking domain of the heterodimeric TadA enzymes. The catalytic

domain of the intermediate TadA7.10 was modelled by introducing its mutations in the ABE8e cryo-EM structure (PDB: 6VPC).[259] Missing residues of the structures were added via homology modelling using AlphaFold2.[278] All systems were embedded in explicit waters, and counterions were added to neutralize the total charge at physiological conditions, leading to periodic cells of $\sim 133.68 \cdot 148.43 \cdot 224.94$ Å^3, for a total of $\sim$500,000 atoms for each system.

E.2 Molecular Dynamics (MD) Simulations

Molecular dynamics (MD) simulations were performed using a simulation protocol tailored for protein/nucleic acid complexes,[279] which we also used in studies of CRISPR-Cas systems.[56,168,280] We employed the Amber ff19SB[126] force field, and included the OL15[281] and OL3[102] corrections for DNA and RNA, respectively. The TIP3P[104] model was used for water molecules. We re-parametrized the Zn^{2+} ion and the active site residues using Quantum Mechanical (B3LYP[282]/6-31G*[283]) calculations, following the protocol by Li & Merz.[284,285]

As a first step, all systems were subjected to energy minimization to overcome the potential inter- and intra-molecular steric clashes. Then, the systems were heated from 0 to 100 K in two consecutive NVT simulations (representing the canonical ensemble) of 5 ps each, imposing positional restraints of 100 kcal mol^{-1} Å^{-2} on the protein, the nucleic acids, and the TadA active site. The temperature was further increased up to 200 K in a subsequent $\sim$100 ps MD run in the isothermal-isobaric ensemble (NPT), in which the restraint was reduced to 25 kcal mol^{-1} Å^{-2}. Finally, all restraints were released, and the

systems were heated up to 300 K in a single NPT simulation of 500 ps. These simulations were performed using a 1 fs time step. The simulation time step was subsequently increased to 2 fs for further equilibration and production simulations. All bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm.[138] After ∼600 ps of equilibration, ∼10 ns of NPT simulation were carried out, allowing the systems' density to stabilize around 1.01 g/cm^3. The temperature was kept constant at 300 K via Langevin dynamics, with a collision frequency γ = 1/ps. The pressure was controlled in the NPT simulations by coupling the system to a Berendsen barostat[108]17 at a reference pressure of 1 atm and with a relaxation time of 2 ps. Finally, production runs were carried out for all systems in the NVT ensemble, reaching ∼3 μs of MD simulations for each system and in three independent replicates (i.e., ∼9 μs of aggregate sampling for each system). The MD simulations replicas were performed starting from different configurations and velocities initialized according to the Maxwell-Boltzmann distribution at physiological temperature. Considering four simulation systems, we collected ∼36 μs of aggregate sampling (i.e., 4 systems * 3 replicas * ∼3 μs = ∼36 μs). These simulations were performed using the GPU-empowered version of AMBER 22.[131]

E.3 Umbrella Sampling (US) Simulations

The Umbrella Sampling (US) method[117] was used to compute the free energy profiles associated with the dimerization of TadA enzymes. In this method, several simulations

(US windows) are run in parallel with additional harmonic bias potential applied to selected Reaction Coordinates (*RC*s):

$$V(RC) = \frac{k}{2}(RC(t) - RC^*)^2 \tag{5.1}$$

where $V(RC)$ is the value of the bias potential, k is a bias force constant, $RC(t)$ is the value of *RC* at given time t and $RC*$ is the reference value of RC. By using different $RC*$ values in each US window, one can sample the biased probability distribution $p_b(RC)$ along the whole *RC* range of interest. Thanks to a reweighting scheme providing unbiased probabilities, the free energy of a process along the RCs can be computed (*vide infra*).

We performed US simulations at 300 K using two RCs. *RC*1 was the inter-monomeric distance considering the center of mass of each TadA unit. *RC*2 was the difference in root-mean-square deviation (RMSD) of all the Cα atoms with respect to the experimental structures (i.e., PDB: 1Z3A[261] for wt-TadA and PDB: 6VPC[259] for TadA8e). The combination of these RCs enabled us to describe the dimerization process while keeping the experimental structures as a reference. Restraint forces were applied on the two RCs as follows. *RC*1 was discretized into bin sizes of 0.5 Å with a collective force constant k of 17.5 kcal mol^{-1} Å^{-2}. *RC*2 was divided into bin sizes of 0.5 Å RMSD difference with a collective force constant k of 25 kcal mol^{-1} Å^{-2}. Each window was run for ∼15 ns. Overall, we simulated 51 (*RC*1) * 31 (*RC*2) = 1,581 windows, resulting in a collective ensemble of ∼25 μs of sampling for each system. Two independent sets of two-dimensional (2D) US simulations were performed considering the wt-TadA and TadA8e.

The free energy profiles were computed using the Variational Free Energy Profile (VFEP) method,[286] which uses the maximum likelihood principle to construct free energy profiles. This method was applied upon removing the initial $1/3^{rd}$ of the US trajectories from each window, considering the system's relaxation. Analysis of the minimum free energy pathway (MFEP) between the dimeric and monomeric states was performed using the MEPSA code,[140] which employs an approach similar to the Dijkstra algorithm to trace the MFEP between nodes of interest. The convergence of the free energy profiles was evaluated by plotting the minimum free energy pathways, considering the remaining 10 ns runs from each US window, for $\sim$5 ns through $\sim$10 ns (Supplementary Figure E.3). The error estimation on the minimum free energy pathways was performed using the Monte Carlo bootstrap error analysis.

E.4 Differential Inter-residue Interactions

To characterize the interactions in the complexes, we employed the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method.[229,265,266] The method was used to obtain the relative change in pairwise interaction energies (ΔE) across different ABE systems. For each system, the pairwise interaction energies E were computed by pairwise decomposition of the MM-GBSA energy. Then, the difference in pairwise interaction energy (ΔE) between any couple of systems (e.g., ABE8e and ABE0.1) was calculated as $\Delta E_{A-B} = E_A - E_B$, where E_A and E_B are the interaction energies in systems A (i.e., ABE8e) and B (ABE0.1), respectively. As interaction energies $E_{A/B}$ are negative in magnitude, $\Delta E_{A-B} < 0$ corresponds to interactions that gain strength in system A,

while $\Delta E_{A-B} > 0$ correspond to interactions relatively more stable in system B. To filter contacts with significant perturbation between the systems, a cut-off of 0.7 kcal/mol was used for reporting E_{A-B} in two-dimensional interaction maps (Supplementary Figs. E.5-E.7). ΔE values were computed between the following systems: ABE8e – ABE0.1, ABE8e – ABE7.10, and ABE8e – ABE8e hetero; and evaluated at three interfaces: (*i*) the TadA catalytic and docking domains, (*ii*) the TadA docking domain and Cas9, (*iii*) the TadA docking domain and the DNA non-target strand (NTS). MM-GBSA calculations were performed using the MMPBSA.py code,[303] available in the AMBER 22[304] suite of codes.

E.5 Alchemical Free Energy Perturbation

We performed alchemical free energy perturbation (FEP) simulations using λ-dynamics,[267] to compute the relative free energy change (i.e., $\Delta\Delta G$) in the ABE7.10 complex while introducing the ABE8e mutations (i.e., A109S, T111R, D119N, H122N, Y147D, F149Y, T166I, D167N). With our alchemical FEP simulations, we aimed to describe the effect of amino acid mutations on the thermal stability of the complex. Hence, we employed an alchemical FEP approach designed to compute the impact of mutations on the folding free energy,[268,269] as thermal stability changes are caused by alterations in the free energy of folding.[268] We followed the thermodynamic cycle in the schematic diagram below.

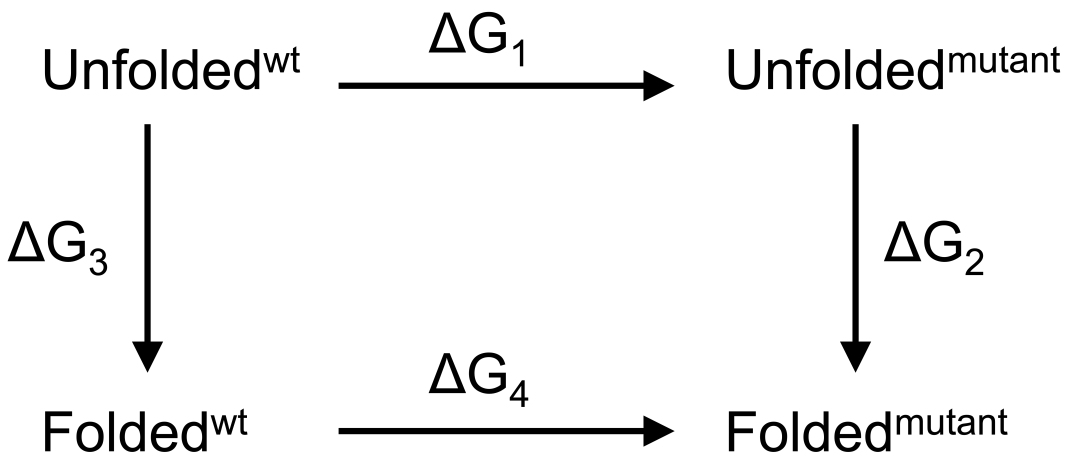


Figure E.1: **FEP Schematic Diagram.** Thermodynamic cycle. The unfolded state was modelled with the AAXAA peptides (X = mutation of interest; A = any flanking amino acid). From the thermodynamic cycle, the folding free energy difference $\Delta\Delta G = \Delta G_3 – \Delta G_2$ between the wild-type protein and the mutant can be calculated via $\Delta G_1 – \Delta G_4$.

The unfolded states have been approximated through a simplified pentapeptide (AAXAA) including the mutation under investigation (i.e., the central mutation, X) and the two amino acid residues (AA) on either side as in the primary sequence. Consistently with a thermodynamic cycle, state *A* was evolved into state *B*, representing the initial and the final residues respectively, via a series of transition states following the alchemical free energy formalism. A lambda (λ) dynamic variable was used to define the transitioning thermodynamic states of the system, where $\lambda = 0$ and $\lambda = 1$ represent the initial and final states respectively. The distribution of the λ parameter was optimized using a gradient descent algorithm to maximize the probabilities of exchange between adjacent states (https://gitlab.com/KomBioMol/converge_lambdas). In detail, the neighbouring λ-windows were allowed to exchange their configurations every 0.5 ps according to the Metropolis criterion, and the values of λ were optimized to achieve the acceptance rate of at least 15%. Since each intermediate λ state is technically a hybrid

between the *A* and *B* end point states, we generated their dual coordinates and topologies using the PMX server (http://pmx.mpibpc.mpg.de/).[288] The relative free energy changes (i.e., $\Delta\Delta G$) for the transformation were computed using the Multistate Bennett Acceptance Ratio (MBAR) method[118,119] to integrate the free energies over the different λ values.[143,144]

Using this method, we computed relative $\Delta\Delta G$ for each single-point mutation introduced through directed evolution in ABE7.10 leading to ABE8e (i.e., A109S, T111R, D119N, H122N, Y147D, F149Y, T166I, D167N). For each mutation, alchemical FEP simulations were performed across the λ values, running $\sim$50 ns for each window and reaching a total of $\sim$8 μs sampling. The convergence of the $\Delta\Delta G$ for each system is shown in Supplementary Figure E.11.

E.6 Cloning and Protein Expression

Recombinant ABE8e* was constructed by fusing two TadA8e domains, each with a unique DNA sequence encoding the same amino acid sequence, to the N terminus of nCas9. The TadA8e domains are linked with a 32 amino acid linker and include a N-terminal His_6-tag. This construct was cloned into bacterial expression plasmid pBR3222 and used as a starting point to generate all other mutants described in Supplementary Table E.1. ABE8e* and its mutants were expressed in Rosetta™ 2(DE3) pLysS competent cells by inducing with 500 mM IPTG for 19 hours at 16 °C, once the cell density reached an OD_{600} of 1.0.

E.7 Protein Purification

Protein was purified as described previously by Lapinaite et. al.[259] Briefly, cells were pelleted by centrifugation, resuspended in the lysis buffer (25 mM HEPES pH 8.0; 1 M KCl; 10 mM imidazole; 2 mM TCEP; 10% glycerol; 0.5% TritonX-100) supplemented with 1 mg/mL lysozyme and protease inhibitors (Roche), and lysed by sonication. The soluble fraction was obtained by centrifugation (20,000 rpm, 1 hour) and then loaded onto a HisTrap FF affinity chromatography column (Cytiva). Several alternating washes were performed using either 10-column volumes of lysis buffer or 5-column volumes of high salt buffer (25 mM HEPES pH 8.0; 1 M KCl; 2 M LiCl; 10 mM imidazole; 2 mM TCEP; 10% (v/v) glycerol). The ABEs were then eluted with a gradient of an elution buffer (25 mM HEPES pH 8.0; 500 mM KCl, 300 mM imidazole; 2 mM TCEP; 10% (v/v) glycerol). Next, the buffer was exchanged to low salt IEX A buffer (25 mM HEPES pH 8.0, 200 mM KCl, 1 mM TCEP and 10% (v/v) glycerol) using a HiPrep 26/10 desalting column (Cytiva). Flowthrough was collected and loaded onto a HiTrap Heparin HP column (Cytiva) to remove nucleic acid contamination. The pure ABEs were eluted from the HiTrap Heparin column with a linear gradient of high salt IEX B buffer (25 mM HEPES pH 8.0; 1.2 M KCl; 1 mM TCEP; 10% glycerol). The eluent was further purified through size exclusion chromatography on a Superdex 200 16/600 column pre-equilibrated with SEC buffer (20 mM Tris-HCl pH 8.0; 200 mM KCl; 10% (v/v) glycerol; 1 mM TCEP). The fractions corresponding to the peak eluting at $\sim$63 mL were collected and concentrated to $\sim$14 mg/mL. Concentrated protein was flash-frozen in liquid nitrogen, and stored at -80 °C. All steps were performed at 4 °C.

E.8 *In vitro* Transcription of Guide RNA

Single guide RNA (sgRNA) was transcribed overnight at 37 °C in buffer containing 30 mM Tris-HCl (pH 8.0), 25 mM $MgCl_2$, 0.01% Triton X-100, 2 mM spermidine-HCl, 5 mM each NTP, 10 mM DTT, 100 μg/mL in-house prepared T7 polymerase, RNase Inhibitor (Promega), thermostable inorganic pyrophosphate (Roche) and 1 μM dsDNA template purchased from IDT. The template contained a consensus 5' T7 promoter sequence as well as sequence complementary to sgRNA. After transcription, 5 units of RNase-free DNase (Promega) were added to the reaction, followed by incubation at 37 °C for 30 minutes. Transcribed RNA was purified by using 7 M urea, 10% 19:1 polyacrylamide gel electrophoresis, followed by crush-and-soak extraction overnight at 4 °C and ethanol precipitation as described previously.[259]

E.9 Assembly of ABE-sgRNA Complexes

Prior to assembling with ABEs, the sgRNAs were diluted in 20 mM HEPES (pH 7.5), 150 mM KCl, 1 mM TCEP, 10% (v/v) glycerol buffer. They were then refolded by incubating at 70 °C for 5 min followed by slow cooling to RT. After reaching RT, $MgCl_2$ was added to a final 1 mM concentration and the sgRNA was incubated for a further 5 min at 50 °C and slowly cooled to RT. The ABE RNP complexes were formed by incubating the protein (Supplementary Table E.1) with the refolded sgRNA at a molar ratio of 1:1.5 (protein to sgRNA) in RNP assembly buffer (20 mM Tris-HCl, 200 mM KCl, 5 mM $MgCl_2$, 5% (v/v) glycerol, 2 mM DTT, pH 7.5) at room temperature for 20 minutes.

E.10 Single-turnover *in vitro* DNA and RNA Deamination Assays

The single turnover DNA deamination reactions were conducted with 2 μM ABE RNPs in a deamination buffer consisting of 20 mM Tris-HCl (pH 7.5), 100 mM KCl, 5% (v/v) glycerol, 2.5 mM $MgSO_4$, and 2 mM DTT. These reactions were initiated by introducing fluorescently labelled dsDNA (the non-target strand was labelled at the 5' end with IRDye® 800CW, purchased from IDT) or hairpin RNA resembling anticodon hairpin of $tRNA^{Arg2}$ (labelled at the 5' end with fluorescein, purchased from IDT) to achieve a final concentration of 7.5 nM. Subsequently, the mixture was incubated at 37 °C. 20 μL aliquots were transferred into 30 μL preheated water at 95 °C at different time intervals (0, 1 min, 2 min, 5 min, 10 min, 20 min and 60 min for ABE8e and ABE0.1 variants and 0, 1 min, 2 min, 5 min, 10 min, 20 min, 60 min, 180 min, 360 min, and 480 min for ABE 7.10 variants) and quenched by heating at 95 °C for 5 min.

For detection of DNA adenine deamination the quenched reactions were then purified using Monarch PCR & DNA Cleanup Kit (NEB). Purified DNA was incubated with 20 units of EndoV (NEB) at 37 °C for 1.5 hours.

For detection of RNA adenine deamination, human EndoV (hEndoV; purified in-house) was added into the reaction at a final concentration of 0.1 μM in cleavage buffer (20 mM Tris-HCl pH 7.5, 50 mM KCl, 5% (v/v) glycerol, 4 mM $MnCl_2$, 2 mM DTT). The reaction was incubated at 37 °C for 1 h.

In both cases, the EndoV cleavage reactions were quenched by adding an equal volume of gel loading buffer supplemented with 15 mM EDTA, 0.1% (w/v) SDS, and

200 μg/mL heparin, and heating at 95 °C for 5 minutes. Samples were resolved on 15% (v/v) denaturing (29:1) polyacrylamide gel. The gel was visualized by using the fluorescence imaging instrument Amersham Typhoon (GE Healthcare). The experiments were conducted in three separate replicates, and the levels of uncleaved and cleaved DNA or RNA were evaluated using ImageQuant TL Software from GE Healthcare.

E.11 Circular Dichroism (CD) Measurements

CD spectra of ABE8e variants were recorded at a concentration of 0.4 mg/mL in the RNP assembly buffer. Measurements were performed using a 1 mm pathlength microcuvette (Hellma 110-1-40, style 110-QS) in the far-UV range (190-240 nm) at room temperature. A single measurement was conducted with an integration time of 0.5 seconds, using a scanning speed of 50 nm/min and a bandwidth of 1 nm on the Jasco J-815 spectropolarimeter. Each spectrum was generated by averaging eight scans while subtracting the scanning spectrum of the RNP buffer, and the ellipticity was recorded in millidegrees (mdeg).

E.12 Thermal Stability Assay using CD Spectroscopy

The sample was prepared and loaded into the 1 mm pathlength microcuvette (Hellma 110-1-40, style 110-QS). The measurements were performed using the Jasco J-815 spectropolarimeter. The ellipticity at the wavelength of 208 nm was recorded from 20 °C to 95 °C at the ramp rate of 1 °C/min. The measurement was performed with an integration time of 2 seconds, using a bandwidth of 1 nm. The ellipticity was plotted as

a function of temperature and fitted with a Boltzmann sigmoidal equation. The melting temperature, T_m, was determined by the inflection point.

E.13 Preparation of Fluorescently-labelled ABEs

Labelling reactions were performed as previously described. Briefly, the labelling reaction was carried out in SEC buffer and contained 10 μM protein along with 200 μM Cy3- and 1.2 mM Cy5-maleimide (APExBIO) (1:6 ratio of donor to acceptor). The dyes were initially dissolved in anhydrous DMSO, ensuring the final concentration of DMSO does not exceed 5% in the labelling reaction. The labelling reaction was initiated by gentle rocking in the dark for 2 hours at room temperature, followed by overnight incubation at 4 °C. Reactions were quenched by adding 10 mM DTT. To separate the free Cy3- and Cy5-maleimide dyes from the labelled ABEs, gel filtration chromatography was performed using a Superdex 200 Increase 10/300 GL column (Cytiva). The purified labelled ABE samples were concentrated, flash-frozen in liquid nitrogen, and stored at -80 °C. For the control labelling reactions, either cysteine-free ABE8e was used, or the labelling was performed using only one of the dyes.

E.14 Supplementary Tables

Table E.1: **ABE Variants.** Variants of ABE used in this study.

#	Protein Name	Construct	Figure
1	ABE0.1	wtTadA–dCas9	E.4c,d,e
2	ABE0.1 (Cy3/Cy5) labeled	wtTadA(T133C/C141S)–dCas9(C277S/C771S)	6.1c; E.4c,d,e
3	ABE8e	TadA8e–dCas9	6.5h; E.4c,f,g; E.14a
4	ABE8e (Cys-free)	TadA8e(C141S/C146S) –dCas9 (C277S/C771S)	E.4b
5	ABE8e (Cy3/Cy5) labeled	TadA8e(T133C/C141S/C146S)–dCas9 (C277S/C771S)	6.1c; E.4b,c,f,g
6	ABE8e*	TadA8e–TadA8e–dCas9	E.9
7	ABE8e*A	dTadA8e(E59A)–TadA8e–dCas9	6.4b,c; E.9; E.10b,d,f
8	ABE8e*B	TadA8e–dTadA8e (E59A)–dCas9	E.9
9	ABE8e*AB	dTadA8e(E59A)–dTadA8e(E59A)–dCas9	E.9

Continued on next page

continued from previous page

#	Protein Name	Construct	Figure
10	ABE8e*A ($R98A^{dock}$)	dTadA8e(E59A/R98A)–TadA8e–dCas9	6.4b; E.10a,b
11	ABE8e*A ($R98A^{cat}$)	dTadA8e(E59A)–TadA8e(R98A)–dCas9	6.4b; E.10a,b
12	ABE8e*A ($R98E^{dock}$)	dTadA8e(E59A/R98E)–TadA8e–dCas9	E.10c,d
13	ABE8e*A ($R98E^{dock}$ /Cas9E1056R)	dTadA8e(E59A/R98E)–TadA8e–dCas9(E1056R)	E.10c,d
14	ABE8e*A ($R129E^{dock}$)	dTadA8e(E59A/R129E)–TadA8e–dCas9	6.4c; E.10e,f
15	ABE8e*A ($R129E^{cat}$)	dTadA8e(E59A)–TadA8e(R129E)–dCas9	6.4c; E.10e,f
16	ABE8e*A ($R129E^{dock}$/ Cas9E1049R)	dTadA8e(E59A/R129E)–TadA8e–dCas9(E1049R)	6.4c; E.10e,f
17	ABE7.10	wtTadA–TadA7.10–dCas9	6.5c; E.12
18	ABE7.10 (T111R)	wtTadA–TadA7.10 (T111R)–dCas9	6.5c; E.12a

Continued on next page

continued from previous page

#	Protein Name	Construct	Figure
19	ABE7.10 (D119R)	wtTadA–TadA7.10 (D119)–dCas9	6.5c; E.12a,b
20	ABE7.10 (T111R/D119R)	wtTadA–TadA7.10 (T111R/D119R)–dCas9	6.5c; E.12a
21	TadA7.10	wtTadA-TadA7.10	6.5b
22	TadA7.10 (T111R)	wtTadA-TadA7.10(T111R)	6.5b
23	TadA7.10 (T111R/D119N)	wtTadA-TadA7.10 (T111R/D119N)	6.5b
24	ABE8e (Cy3/Cy5) labeled (N119D/N122H)	TadA8e(N119D/N122H/ T133C/C141S/C146S)– dCas9 (C277S/C771S)	6.5g; E.14a,b,c
25	ABE8e (N119D/N122H)	TadA8e(N119D/N122H)- dCas9	6.5h; E.14a,b,c

E.15 Supplementary Figures

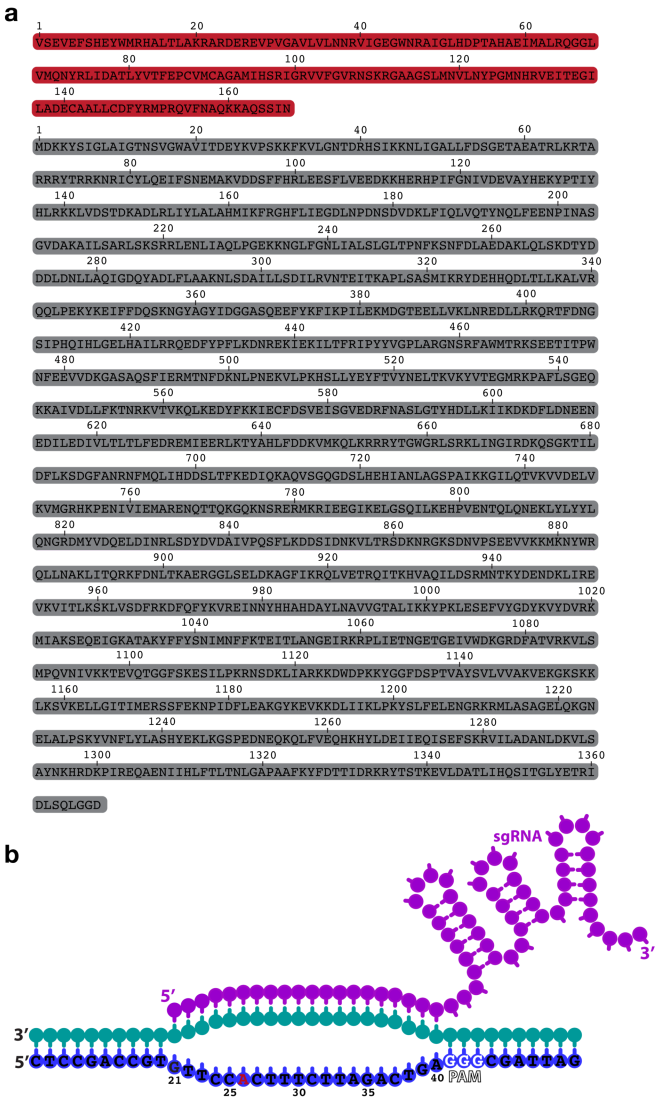


Figure E.2: **Nomenclature of sequences used in this study.** **(a)** Sequences and numbering of TadA8e domain (based on PDB 6V1) and dCas9 (based on PDB 5F9R.[305] **(b)** Schematic diagram of guide and target nucleic acids. The single guide RNA (sgRNA) is in purple, the TS is cyan and the NTS is blue. The Protospacer Adjacent Motif (PAM) recognition sequence is also indicated. The numbering of the NTS single-stranded region was used to denote the nucleobases (based on 6VPC). Target adenine 26 is coloured in red.

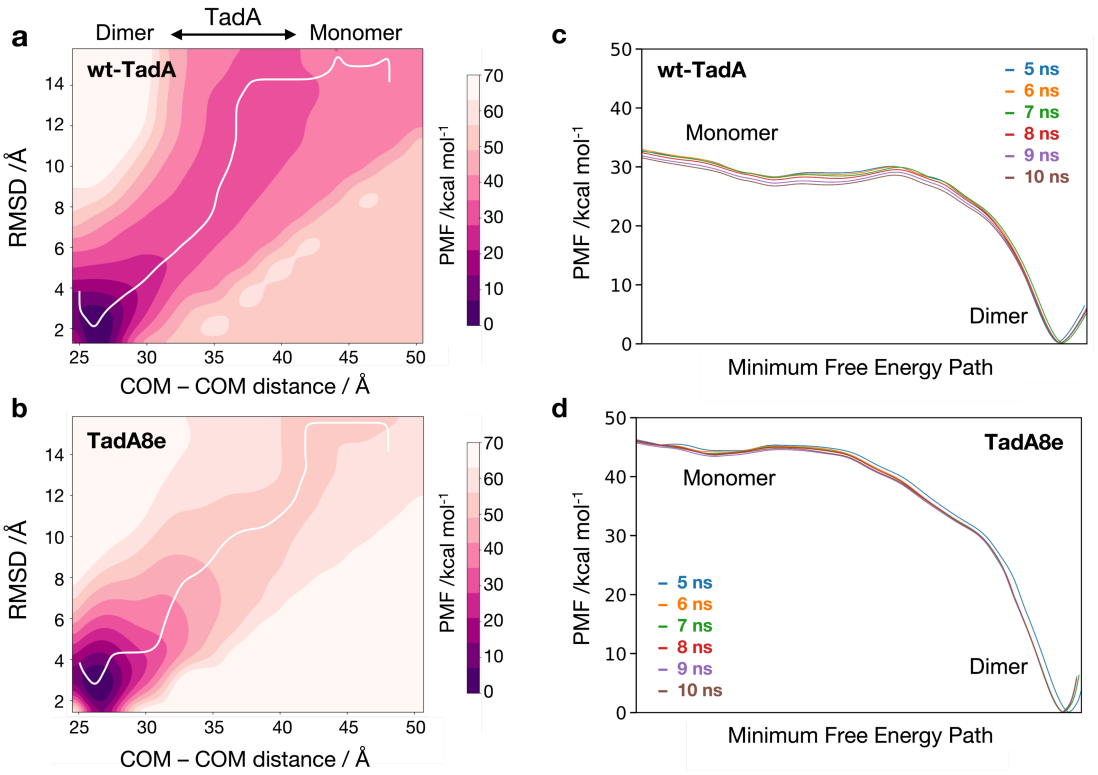


Figure E.3: **Free energy profiles obtained through Umbrella Sampling (US) simulations of the wt-TadA and TadA8e. (a - b)** Two-dimensional free energy surfaces for the dimerization of the wt-TadA **(a)** and TadA8e **(b)** obtained from US simulations along (*i*) the inter-monomeric distance considering the centre of mass of each TadA unit, and (*ii*) the root-mean-square deviation of the Cα atoms between the monomeric and dimeric states of TadA. The dimer and monomer states of TadA enzymes are indicated and shown on top of the plots. The Minimum Free Energy path for the transition of the wt-TadA and TadA8e enzymes from monomer to dimer is shown using a white line (details in the Supplementary Methods). **(c - d)** Convergence of the potential of mean force (PMF) along the minimum free energy path for US simulations of the wt-TadA **(c)** and TadA8e **(d)**. The convergence of the free energy profiles was evaluated by plotting the minimum free energy pathways, considering the remaining ∼10 ns runs from each US window, for ∼5 ns through ∼10 ns. The PMF (kcal/mol) was computed for sampling until 5 ns, 6 ns, 7 ns, 8 ns, 9 ns and 10 ns after removing the biases incurred by the restraints during the sampling. Convergence is reached ∼8 ns onward for both the systems and we used bin sizes of ∼10 ns to compute our 2-D free energy surfaces and the minimum free energy paths (Figure 6.2, main text).

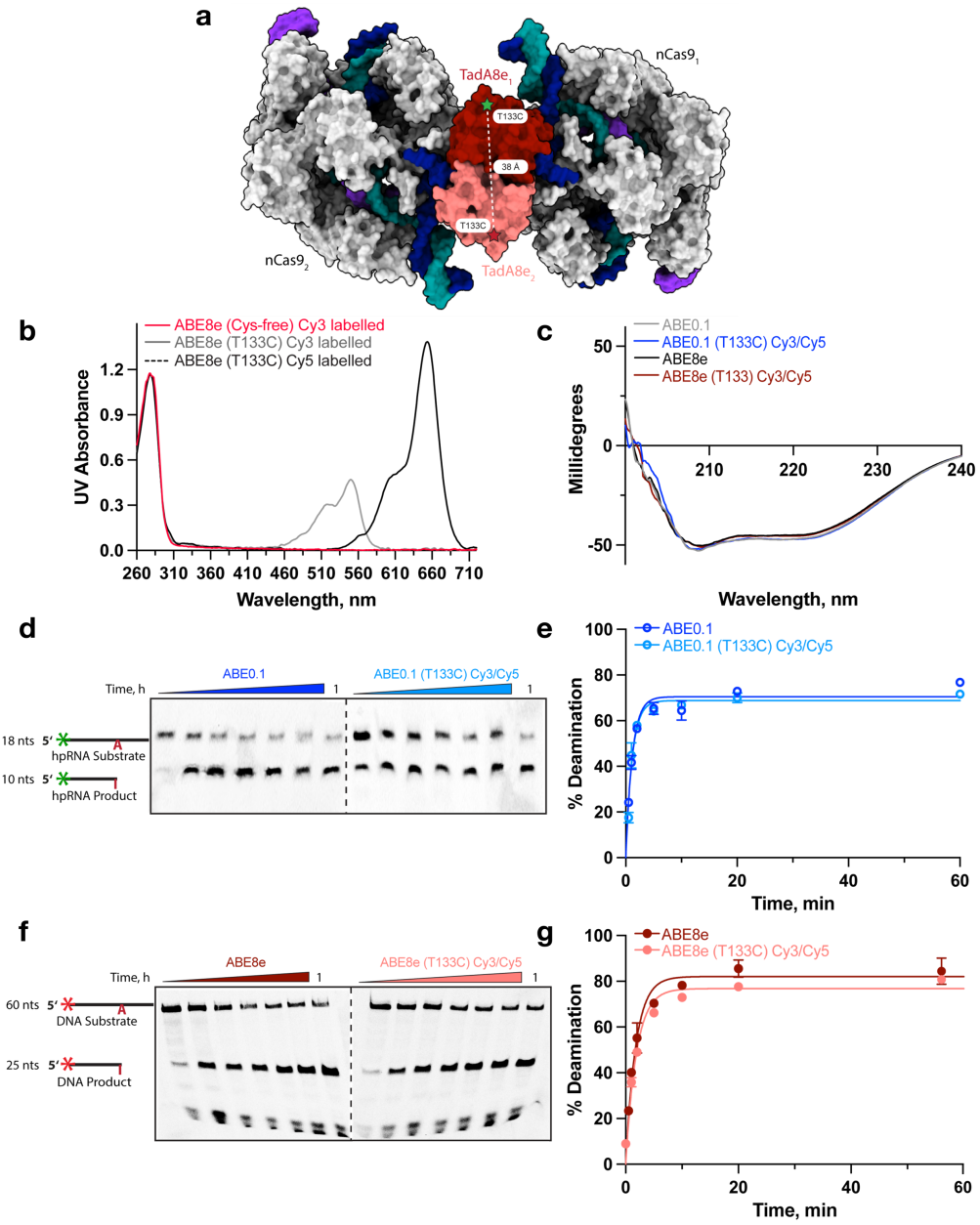

a
TadA8e1
nCas91
T133C
38 Å
T133C
nCas92
TadA8e2
b
ABE8e (Cys-free) Cy3 labelled
ABE8e (T133C) Cy3 labelled
ABE8e (T133C) Cy5 labelled
UV Absorbance
Wavelength, nm
c
ABE0.1
ABE0.1 (T133C) Cy3/Cy5
ABE8e
ABE8e (T133) Cy3/Cy5
Millidegrees
Wavelength, nm
d
Time, h
ABE0.1
ABE0.1 (T133C) Cy3/Cy5
18 nts 5′
hpRNA Substrate
10 nts 5′
hpRNA Product
e
ABE0.1
ABE0.1 (T133C) Cy3/Cy5
% Deamination
Time, min
f
Time, h
ABE8e
ABE8e (T133C) Cy3/Cy5
60 nts 5′
DNA Substrate
25 nts 5′
DNA Product
g
ABE8e
ABE8e (T133C) Cy3/Cy5
% Deamination
Time, min

Figure E.4: **Design of ABE FRET constructs and biochemical and biophysical characterization of dye-labelled ABE0.1 and ABE8e.** **(a)** Design of ABE FRET constructs. The measured distance between T133 in wt-TadA (PDB 1Z3A) and TadA8e (PDB 6VPC) dimers is expected to be $\sim$38 Å. This distance can be detected by the presence of acceptor fluorescence emission (a peak at 675 nm for Cy5) when exciting only the donor (at 510 nm for Cy3). **(b)** UV-visible spectra of ABE8e variants: the cysteine-free ABE8e (red) after the labelling with Cy3 dye, ABE8e (T133C) labelled with Cy3 (grey), ABE8e (T133C) labelled with Cy5 (black). The absence of absorption peaks between 460 and 610 nm wavelengths in the cysteine-free ABE8e after labeling suggests that this construct does not undergo labeling and is suitable for the preparation of FRET constructs for site-specific labeling. **(c)** Far UV (below 250 nm) CD spectra for ABE0.1 (grey), Cy3/Cy5-labled ABE0.1 (T133C) (blue), ABE8e (black) and Cy3/Cy5-labelled ABE8e (T133C) (red). All four proteins acquire similar $\alpha + \beta$ fold. **(d)** A representative gel of single turnover in vitro RNA deamination by ABE0.1 (left) and dye-labelled ABE0.1 (T133C) (right) using fluorescently labeled hairpin RNA, resembling tRNA anticodon loop, as a substrate. Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 min. Concentrations of RNPs were 2 μM and concentrations of RNA were 7.5 nM for single-turnover assays. **(e)** Single-turnover kinetics of ABE0.1 and dye-labelled ABE0.1 (T133C) variant using hpRNA containing a single adenosine within the editing window. The fraction (in %) of deaminated RNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments. The apparent deamination rates (k_{obs}) are: ABE0.1 0.87 $\pm$ 0.06 min^{-1}; and dye-labelled ABE0.1 (T133C) 0.85 $\pm$ 0.06 min^{-1}. **(f)** A representative gel of single turnover in vitro dsDNA deamination by ABE8e (left) and dye-labelled ABE8e (T133C) (right). Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 min. Concentrations of RNPs were 1 μM and concentrations of DNA were 7.5 nM for single-turnover assays. **(g)** Single-turnover kinetics of ABE8e and dye-labelled ABE8e variant using dsDNA containing a single adenosine within the editing window. The fraction (in %) of deaminated DNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments. The apparent deamination rates are: ABE8e 0.56 $\pm$ 0.06 min^{-1}; and dye-labelled ABE8e 0.53 $\pm$ 0.05 min^{-1}.

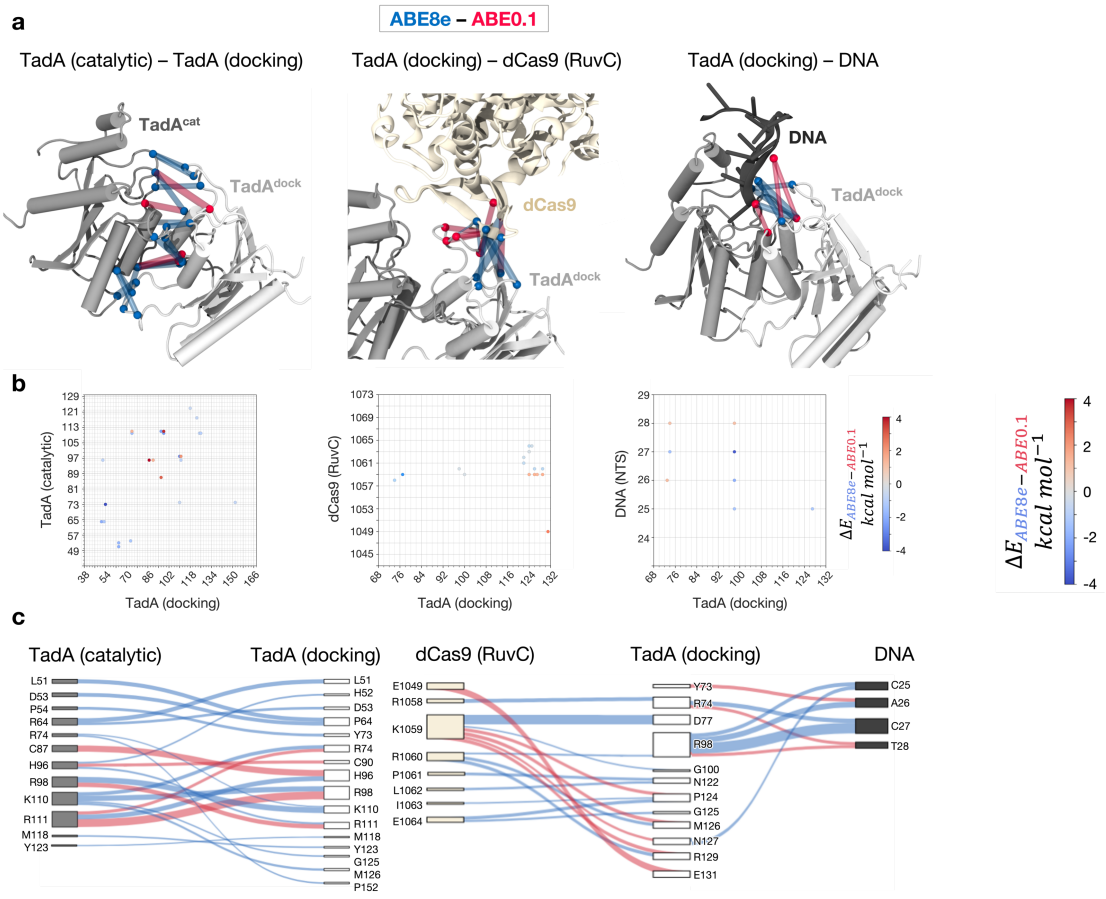

a
ABE8e – ABE0.1
TadA (catalytic) – TadA (docking)
TadA (docking) – dCas9 (RuvC)
TadA (docking) – DNA
TadA^cat
TadA^dock
dCas9
TadA^dock
DNA
TadA^dock
b
TadA (catalytic)
TadA (docking)
dCas9 (RuvC)
TadA (docking)
DNA (NTS)
TadA (docking)
ΔE_ABE8e−ABE0.1 kcal mol^−1
c
TadA (catalytic)
TadA (docking)
dCas9 (RuvC)
TadA (docking)
DNA
L51
D53
P54
R64
R74
C87
H96
R98
K110
R111
M118
Y123
L51
H52
D53
P64
Y73
R74
C90
H96
R98
K110
R111
M118
Y123
G125
M126
P152
E1049
R1058
K1059
R1060
P1061
L1062
I1063
E1064
Y73
R74
D77
R98
G100
N122
P124
G125
M126
N127
R129
E131
C25
A26
C27
T28

Figure E.5: **Gain/loss of interaction strength at the Tada–Cas9–DNA triad of interfaces between ABE8e and ABE0.1.** **(a)** The relative pairwise interaction energies (ΔE) between ABE8e (blue) and ABE0.1 (red) are plotted on the three-dimensional structure of the complex, and shown at three interfaces: the TadA catalytic and docking domains (left), the TadA docking domain and the RuvC domain of nCas9 (centre), the TadA docking domain and the DNA non-target strand (right). Interactions that gain strength in ABE8e (blue) are compared to those that gain strength in ABE0.1 (red). Bonds are proportional to ΔE (as from the 2D maps in panel b). **(b)** Two-dimensional maps of the differential pairwise interaction energies (ΔE) between ABE8e and ABE0.1. ΔE values were computed from residue-pair decomposition of the MM/GBSA energy, as $\Delta E_{A-B} = E_A - E_B$, where E_A and E_B are the interaction energies in systems A (i.e., ABE8e) and B (ABE0.1), respectively. As interaction energies $E_{A/B}$ are negative in magnitude, $\Delta E_{A-B} < 0$ corresponds to interactions that gain strength in system A (ABE8e), while $\Delta E_{A-B} > 0$ correspond to interactions relatively more stable in system B (ABE0.1). ΔE_{A-B} is expressed in kcal/mol. A cut-off of 0.7 kcal/mol was used to filter interactions with substantial perturbation between the systems (details in the Supplementary Methods). **(c)** Sankey diagrams of the relative change in pairwise interaction energy (ΔE) between ABE8e and ABE0.1, computed at the interfaces between the TadA catalytic and docking domains (left), the TadA docking domain, the Cas9 RuvC domain, and the DNA non-target strand (right). Sankey diagrams connect residues through lines the thickness of which is proportional to ΔE (as from panel b). A blue/red colour code is used to indicate the gain of interaction strength in ABE8e ($\Delta E < 0$, blue) and ABE0.1 ($\Delta E > 0$, red). Details in the Supplementary Methods.

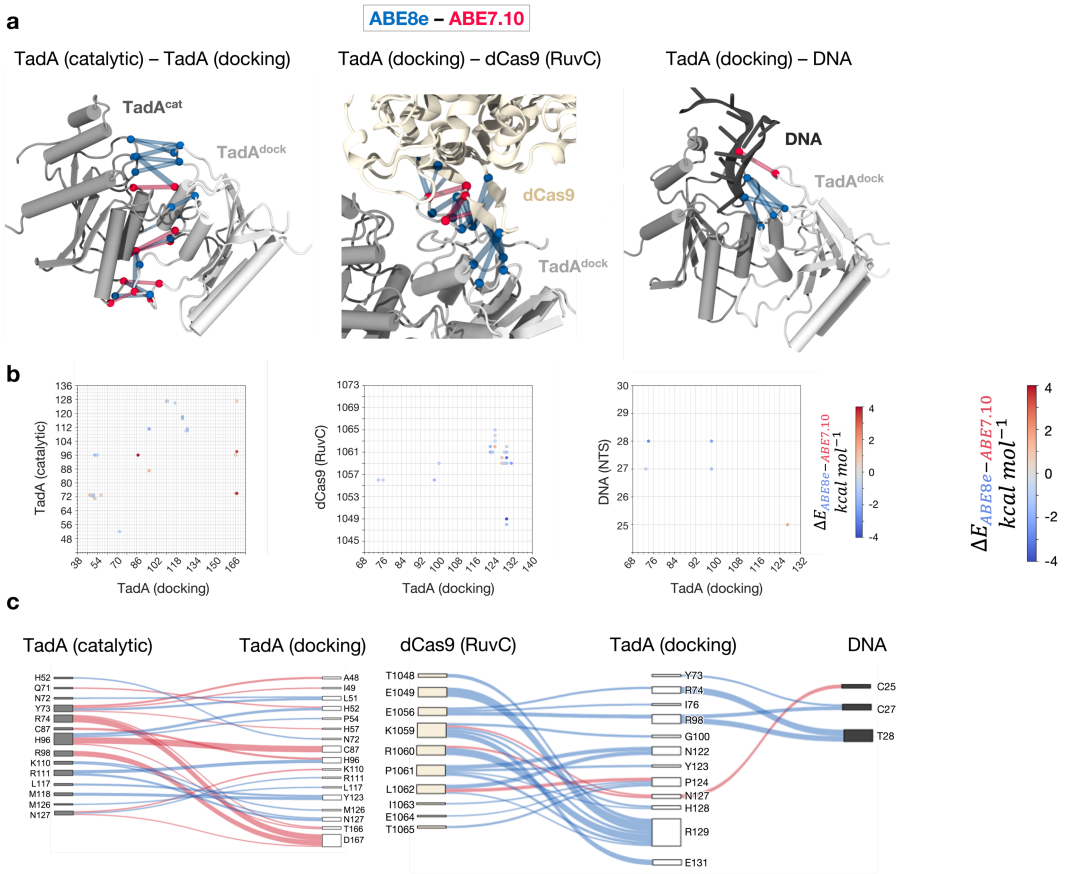


Figure E.6: **Gain/loss of interaction strength at the Tada–Cas9–DNA triad of interfaces between ABE8e and ABE7.10. (a)** The relative pairwise interaction energies (ΔE) between ABE8e (blue) and ABE7.10 (red) are plotted on the three-dimensional structure of the complex, and shown at three interfaces: the TadA catalytic and docking domains (left), the TadA docking domain and the RuvC domain of nCas9 (centre), the TadA docking domain and the DNA non-target strand (right). Interactions that gain strength in ABE8e (blue) are compared to those that gain strength in ABE7.10 (red). Bonds are proportional to ΔE (as from the 2D maps in panel b). **(b-c)** Two-dimensional maps (b) and Sankey plots (c) of the differential pairwise interaction energies computed from residue-pair decomposition of the MM/GBSA energy, as $\Delta E_{A-B} = E_A - E_B$, where E_A and E_B are the interaction energies in systems A (i.e., ABE8e) and B (ABE7.10). Interactions that gain strength in ABE8e ($\Delta E < 0$, blue) are compared to those that gain strength in ABE7.10 ($\Delta E > 0$, red). ΔE_{A-B} is expressed in kcal/mol. Lines in Sankey diagrams are proportional to ΔE. Details are in the caption of Supplementary Figure E.5 and in the Supplementary Methods.

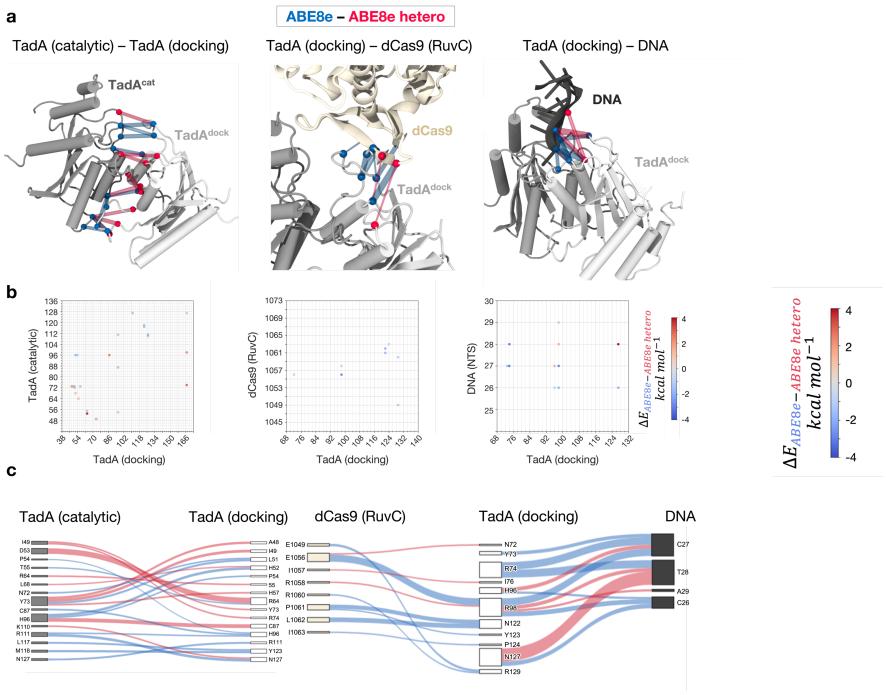


Figure E.7: **Gain/loss of interaction strength at the Tada–Cas9–DNA triad of interfaces between ABE8e and ABE8e including a heterodimeric TadA8e (ABE8e hetero). (a)** The relative pairwise interaction energies (ΔE) between ABE8e (blue) and ABE8e hetero (red) are plotted on the 3D structure of the complex, and shown at three interfaces: the TadA catalytic and docking domains (left), the TadA docking domain and the RuvC domain of nCas9 (centre), the TadA docking domain and the DNA non-target strand (right). Interactions that gain strength in ABE8e (blue) are compared to those that gain strength in ABE8e hetero (red). Bonds are proportional to ΔE (as from the 2D maps in panel b). **(b-c)** 2D maps (b) and Sankey plots (c) of the ΔE computed from residue-pair decomposition of the MM/GBSA energy, as $\Delta E_{A-B} = E_A - E_B$, where E_A and E_B are the interaction energies in systems A (i.e., ABE8e) and B (ABE8e hetero). Interactions that gain strength in ABE8e ($\Delta E < 0$, blue) are compared to those that gain strength in ABE7.10 ($\Delta E > 0$, red). ΔE_{A-B} is expressed in kcal/mol. Lines in Sankey diagrams are proportional to ΔE. Details are in the caption of Supplementary Figure E.5 and in the Supplementary Methods.

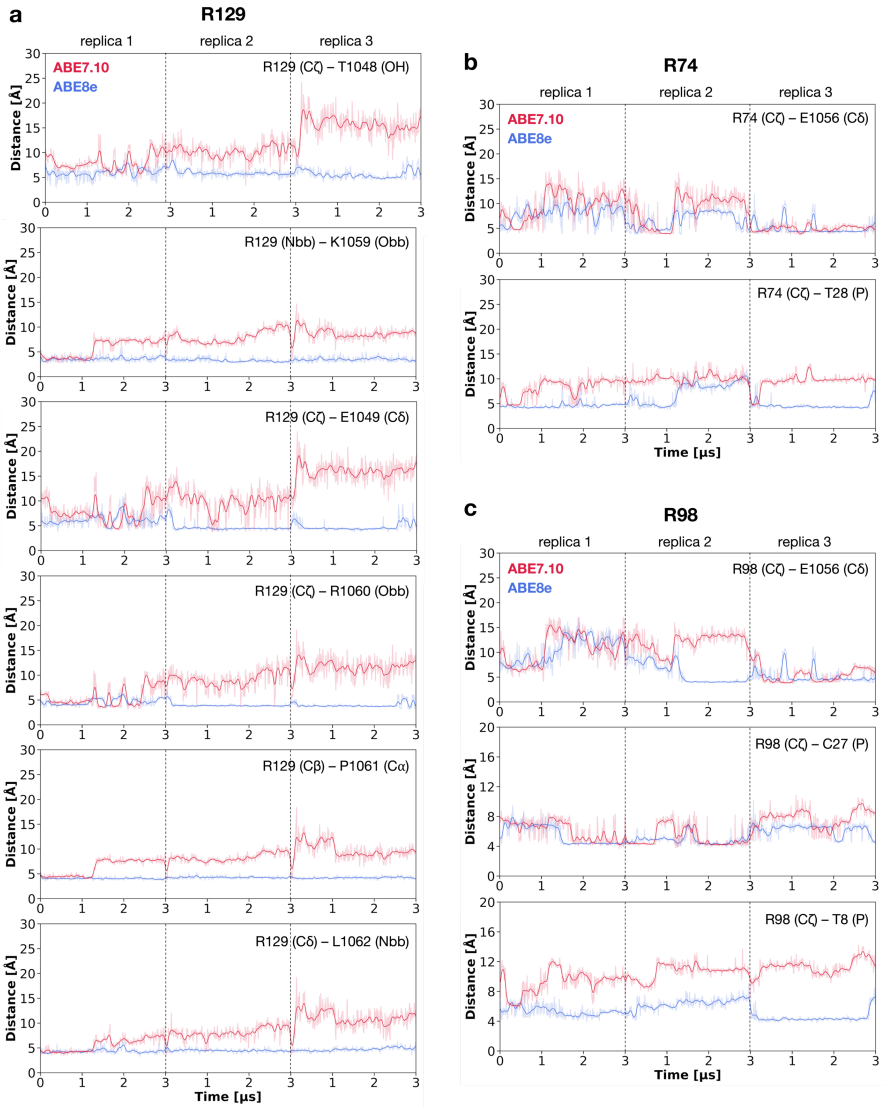


Figure E.8: **Critical interactions of the TadA docking unit in ABE8e and ABE7.10.** Time-evolution of the interactions established by the R129 **(a)**, R74 **(b)** and R98 **(c)** residues of the TadA docking unit with residues of the RuvC domain of nCas9 and with the DNA non-target strand (NTS) along molecular dynamics simulations of the ABE7.10 (red) and ABE8e (blue) complexes. Data are reported for three simulation replicas of $\sim$3 μs each. Salt-bridge interactions are measured considering the Cζ atom of arginine residues (centre of the guanidinium group) and the Cδ atom of glutamate residues (centre of the carboxyl group). Hydrophobic interactions are assessed by measuring the proximity of residues.

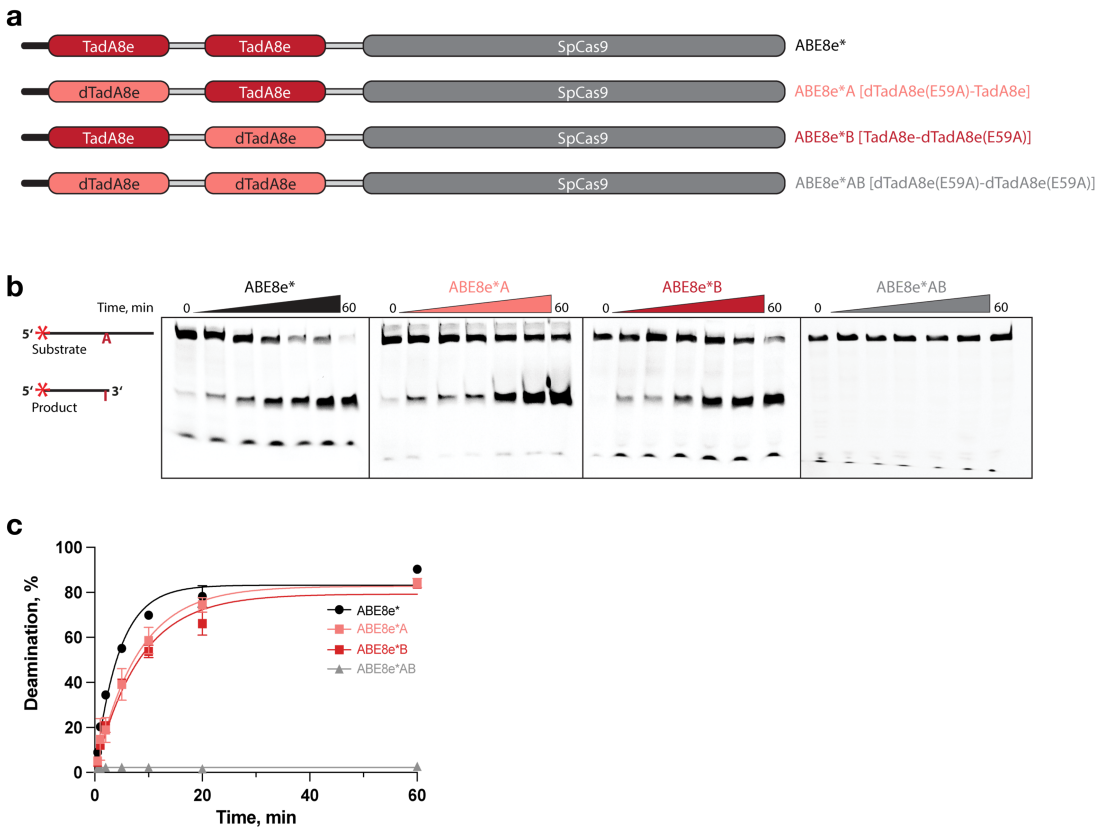


Figure E.9: **Design of constructs for evaluation of TadA8e docking domain's impact on DNA adenine deamination efficiency.** **(a)** Architecture of four ABE8e constructs each containing two TadA8e domains fused to a single Cas9 protein. In all cases, the DNA sequences encoding each TAdA8e within a construct are different, while amino acid sequences remain the same. Construct ABE8e*A has the N-terminal TadA8e inactivated, construct ABE8e*B has the C-terminal TadA8e inactivated, construct ABE8e*AB has both TadA8e domains inactivated via the E59A mutation. **(b)** Gels representing single-turnover kinetics of all four ABE8e constructs (shown in panel a) measured using dsDNA containing single adenine within the editing window. Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 min. Concentrations of RNPs were 1 μM and concentrations of DNA were 7.5 nM for single-turnover assays. **(c)** Single-turnover kinetics of ABE8e RNP constructs (shown in panel a) measured using fluorescently labelled dsDNA containing a single adenine within the editing window. The fraction (in %) of deaminated DNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments. The observed rates of DNA deamination by constructs ABE8e*A (k_{obs} = 0.12 $\pm$ 0.01 min^{-1}) and ABE8e*B (k_{obs} = 0.12 $\pm$ 0.01 min^{-1}) are identical and are half of the rate observed for ABE8e* (k_{obs} = 0.24 $\pm$ 0.01 min^{-1}) when both TadA8e units were active. DNA deamination was not observed for ABE8eAB due to the inactivation of both TadA8e units.

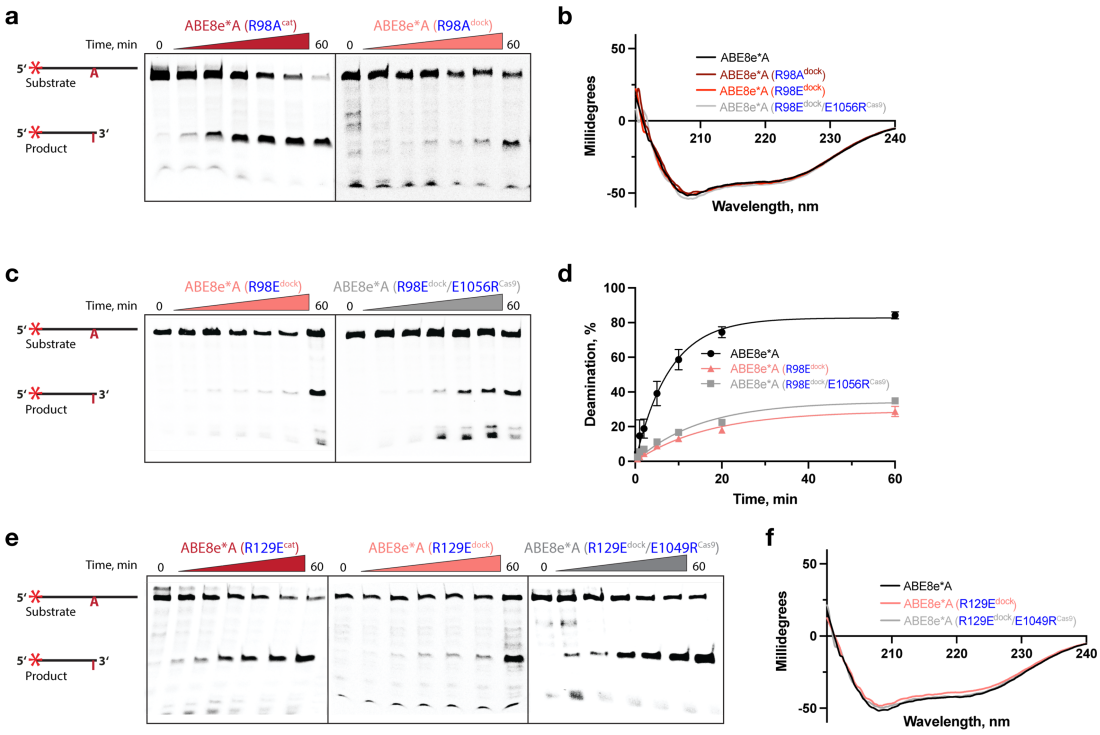

a
ABE8e*A (R98A^cat)
ABE8e*A (R98A^dock)
Time, min
0
60
5′
A
Substrate
Product
3′
b
ABE8e*A
ABE8e*A (R98A^dock)
ABE8e*A (R98E^dock)
ABE8e*A (R98E^dock/E1056R^Cas9)
Millidegrees
Wavelength, nm
c
ABE8e*A (R98E^dock)
ABE8e*A (R98E^dock/E1056R^Cas9)
d
Deamination, %
Time, min
e
ABE8e*A (R129E^cat)
ABE8e*A (R129E^dock)
ABE8e*A (R129E^dock/E1049R^Cas9)
f

Figure E.10: **The impact of TadA8e docking domain on DNA adenine deamination efficiency.** **(a)** Gels representing single-turnover kinetics of ABE8e*A variants with the R98A substation in either the catalytic TadA8e (left, red) or the docking TadA8e (right, pink) domain performed using dsDNA containing single adenine within the editing window. Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 min. Concentrations of RNPs were 1 μM and concentrations of DNA were 7.5 nM for single-turnover assays. **(b)** Far UV (below 250 nm) CD spectra for ABE8e*A (black), ABE8e*A (R98A^{dock}) (red), ABE8e*A (R98E^{dock}) (pink), and ABE8e*A (R98E^{dock}/E1056R^{Cas9}) (grey). All four proteins acquire similar $\alpha + \beta$ fold. **(c)** Gels representing single-turnover kinetics of ABE8e*A variants with the R98E substation in docking TadA8e and either in the absence (left, pink) or presence (right, grey) of the rescue substitution in Cas9 E1056R. Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 min. Concentrations of RNPs were 1 μM and concentrations of DNA were 7.5 nM for single-turnover assays. **(d)** Single-turnover kinetics of ABE8e*A (black) and its variants with the R98E substation in docking TadA8e and either in the absence (pink) or presence (grey) of the rescue substitution in Cas9 E1056R. The fraction (in %) of deaminated DNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments. The DNA deamination rate (k_{obs} = 0.01 $\pm$ 0.01 min^{-1}) and efficiency of the ABE8e*A (R98E^{dock}) are significantly reduced. Introduction of the rescue mutation to the Cas9 (ABE8e*A (R98E^{dock}/E1056R^{Cas9})) does not increases neither the rate (k_{obs} = 0.01 $\pm$ 0.01 min^{-1}) nor the efficiency of DNA deamination to the levels of ABE8e*A. **(e)** Gels representing single-turnover kinetics of ABE8e*A variants with the R129E substation in either the catalytic TadA8e (left, red) or the docking TadA8e (middle, pink). Third gel (grey) represents single-turnover kinetics of ABE8e*A (R129Edock) rescue mutant with E1049R mutation in the Cas9. Assays were performed using dsDNA with single adenine within the editing window in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 min. Concentrations of RNPs were 1 μM and concentrations of DNA were 7.5 nM for single-turnover assays. **(f)** Far UV (below 250 nm) CD spectra for ABE8e*A (black), ABE8e*A (R129E^{dock}) (pink), and ABE8e*A (R129E^{dock}/E1049R^{Cas9}) (grey). All three proteins acquire similar $\alpha + \beta$ fold.

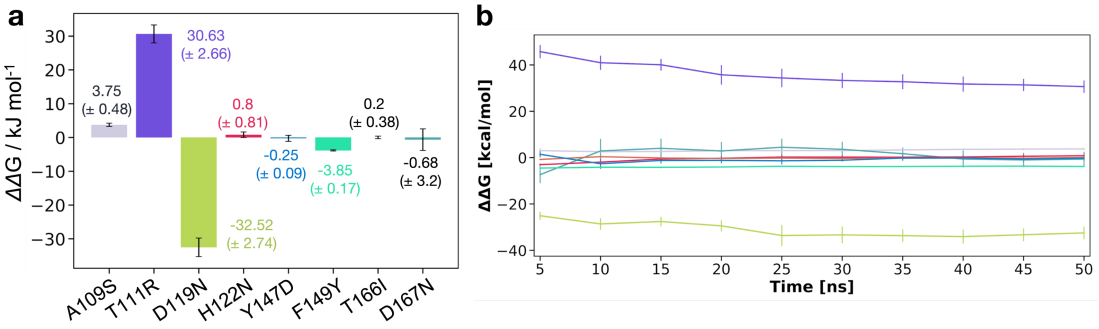


Figure E.11: **Convergence of the relative free energy changes obtained through alchemical free energy perturbation (FEP) simulations.** **(a)** Relative free energy changes ($\Delta\Delta G$, in kJ/mol) in the stability of the ABE7.10 complex while introducing the ABE8e mutations (i.e., A109S, T111R, D119N, H122N, Y147D, F149Y, T166I, D167N). Binding free energies are computed through the Multistate Bennett Acceptance Ratio (MBAR)[118,119] method (see the Methods section) and denoted with the associated MBAR error. **(b)** Convergence profile of the relative binding free energies computed in steps of 5 ns from 5 ns to 50 ns of the alchemical free energy simulations are shown with the MBAR error. Convergence is reached ∼40 ns onward.

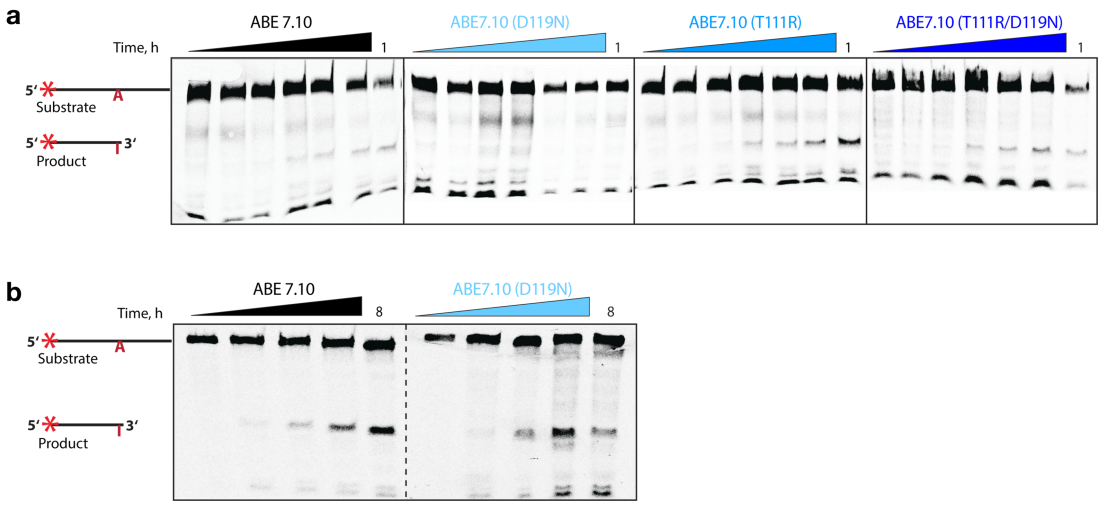


Figure E.12: **Denaturing polyacrylamide gels representing single-turnover kinetics of ABE7.10 variants.** **(a)** Denaturing polyacrylamide gels representing single-turnover kinetics of ABE7.10 (black) and ABE7.10 variants with the substitutions of D119N (light blue), T111R (blue) and both T111R and D119N (dark blue) in TadA7.10 domain performed using dsDNA containing single adenine within the editing window. Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 minutes. **(b)** Denaturing polyacrylamide gels representing single-turnover kinetics of ABE7.10 (black) and its mutant with D119N in TadA7.10 domain (light blue). Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 3, and 8 hours.

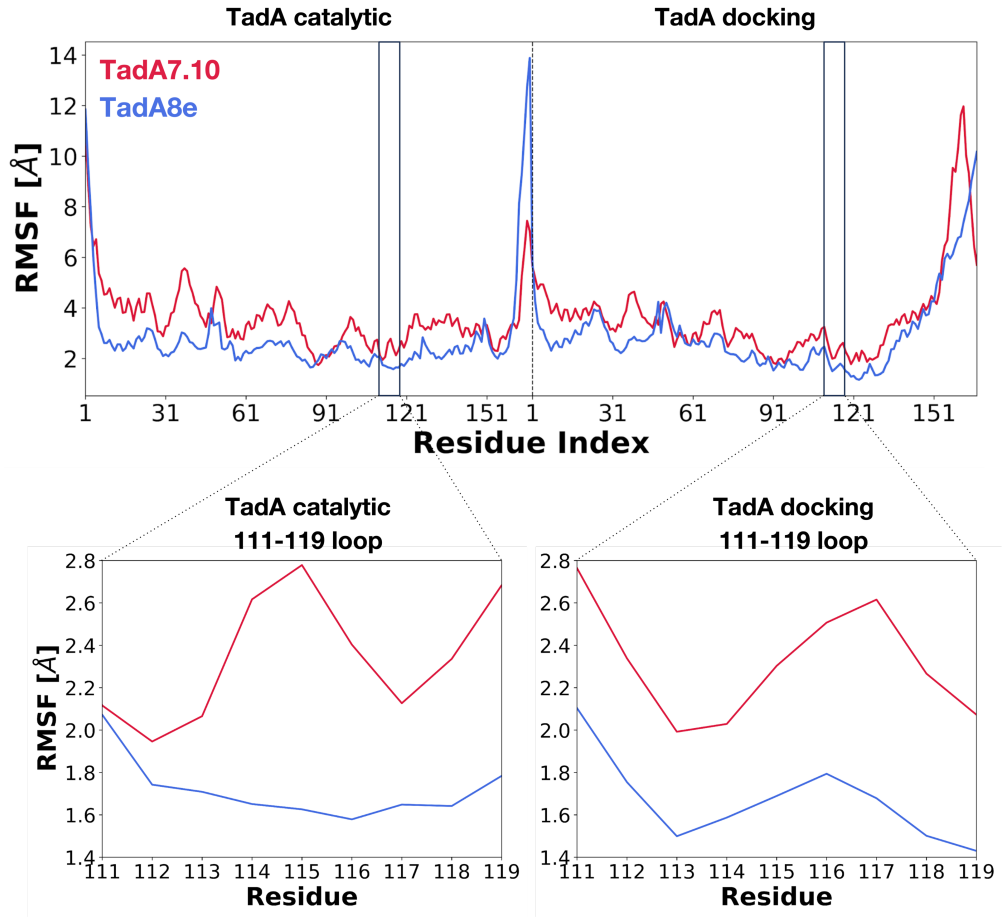


Figure E.13: **Flexibility of TadA7.10 and TadA8e from molecular dynamics simulations of ABE7.10 and ABE8e, respectively.** Root-Mean-Square Fluctuations (RMSF) of the Cα atoms of the TadA7.10 (red) and TadA8e (blue) deaminating enzymes with a close-up view of the region comprising the 111-199 loop (bottom, shown for the TadA catalytic (left) and docking (right) domains). For each system, the aggregate sampling of three simulation replicates of ∼3 μs each was considered.

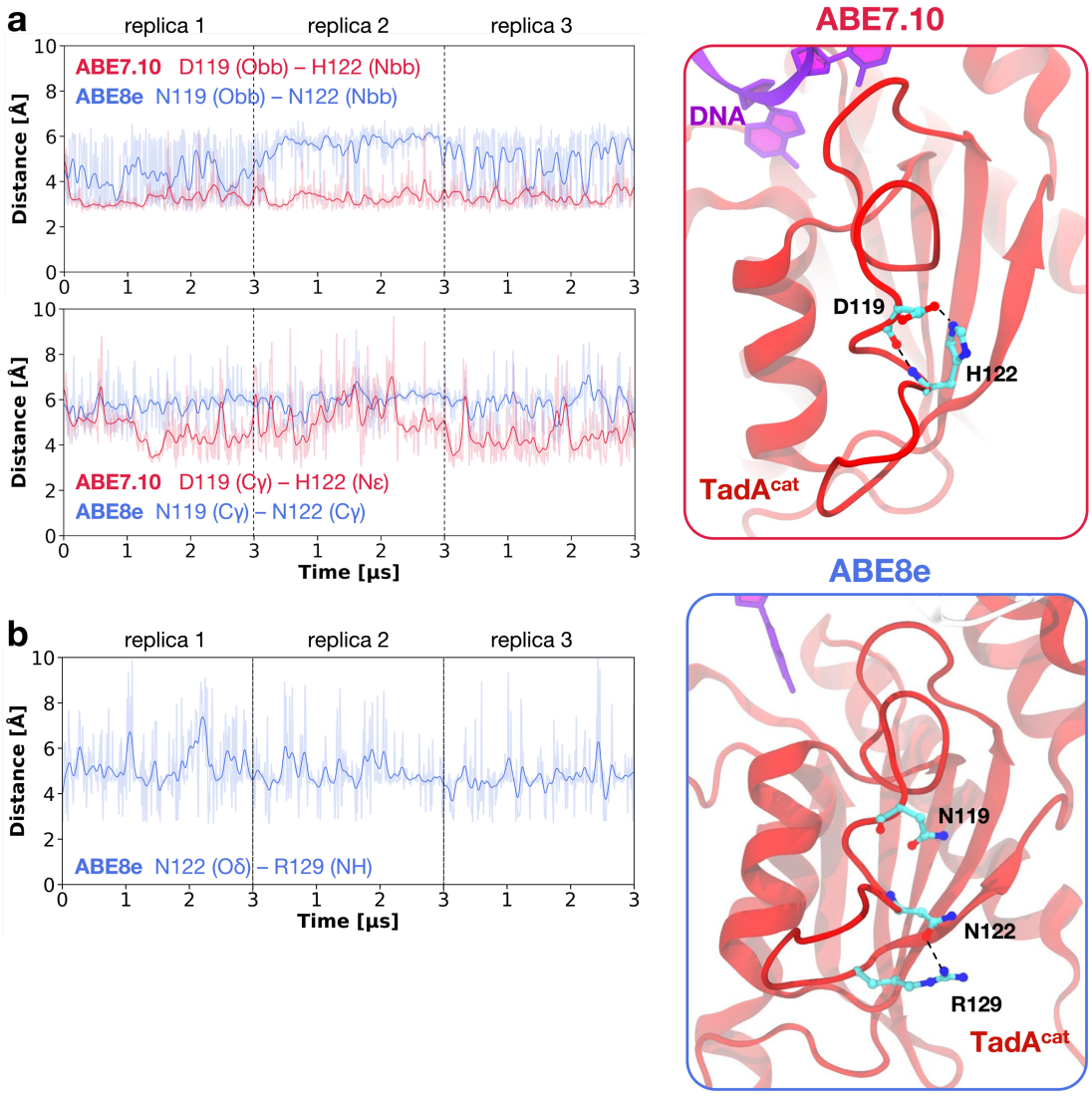


Figure E.14: **Intra-domain interactions established by the 111-119 loop in ABE7.10 and ABE8e.** Time-evolution of the interactions between the D119N and H122 residues of the catalytic domain in ABE7.10 (red) and ABE8e (blue) **(a)**, and between N122 and R129 in ABE8e **(b)** along molecular dynamics simulations of the ABE complexes. Data are reported for three simulation replicates of $\sim$3 μs each. Interactions are measured considering the atoms indicated in parenthesis. The figures on the right show the interactions in ABE7.10 **(a)** and ABE8e **(b)**.

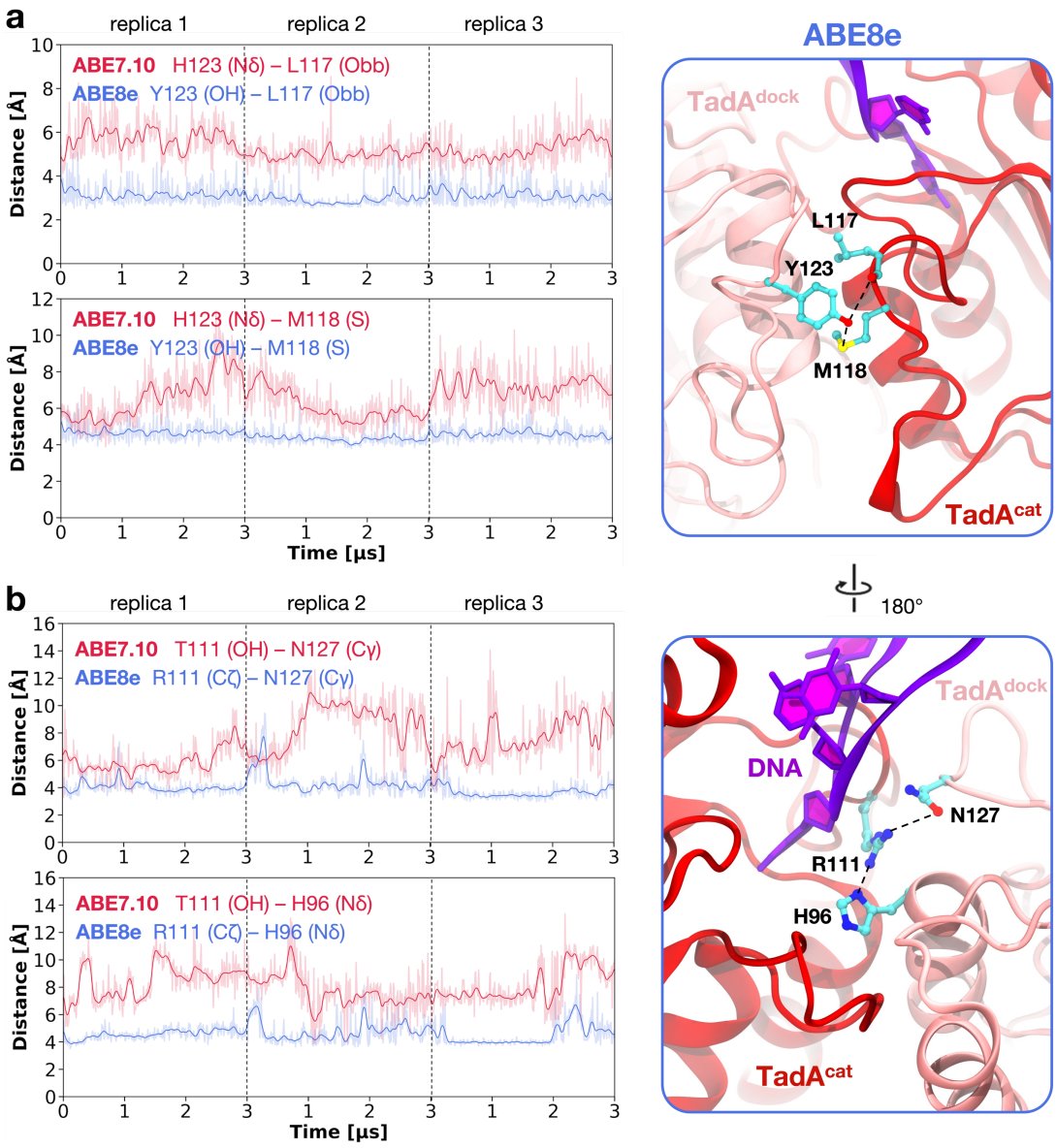


Figure E.15: **Inter-domain interactions established by the 111-119 loop in ABE7.10 and ABE8e.** Time-evolution of the interactions established by the 111-119 loop residues of the catalytic domain with the docking domain residues along molecular dynamics simulations of the ABE8e (red) and ABE7.10 (blue) complexes. H123Y of the catalytic domain interacts with L117 and M118 of the docking domain **(a)**. T111R of the catalytic domain interacts with N127 and H96 the docking domain **(b)**. Data are reported for three simulation replicates of $\sim$3 μs each. Interactions are measured considering the atoms indicated in parenthesis. The figures on the right show the interactions in ABE8e.

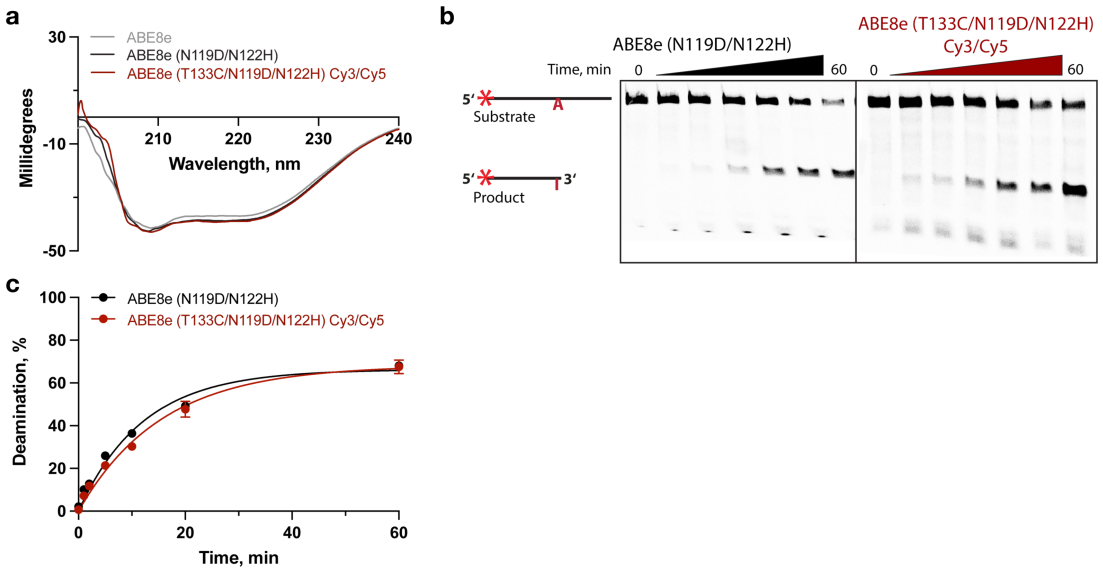


Figure E.16: **Biophysical and biochemical characterization of dye-labelled ABE8e (T133C/N119D/N122H). (a)** Far UV (below 250 nm) CD spectra for ABE8e (grey), ABE8e (N119D/N122H) (black0 and Cy3/Cy5-labelled ABE8e (T133C/N119D/N122H) (red). All three proteins acquire similar $\alpha+\beta$ fold. **(b)** A representative gel of single turnover in vitro dsDNA deamination by ABE8e (N119D/N122H) (left) and dye-labelled ABE8e (T133C/N119D/N122H) (right). Assays were performed in three independent replicates, and time points were taken at 0, 1, 2, 5, 10, 20, and 60 min. Concentrations of RNPs were 1 μM and concentrations of DNA were 7.5 nM for single-turnover assays. **(c)** Single-turnover kinetics of ABE8e (N119D/N122H) and dye-labelled ABE8e (T133C/N119D/N122H) variants using dsDNA containing a single adenosine within the editing window. The fraction (in %) of deaminated DNA was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean $\pm$ SD from three independent experiments. The apparent deamination rates are: Cy3/Cy5-labelled ABE8e (T133C/N119D/N122H) 0.064 $\pm$ 0.004 min^{-1}; and ABE8e (N119D/N122H) 0.08 $\pm$ 0.01 min^{-1}.

E.16 Statistics and Reproducibility

Molecular dynamics simulations were performed in replicates, as also specified in the main text and pertinent figure legends. For all single-turnover kinetics assays, the fraction (in %) of deaminated substrate was plotted as a function of time and fitted to a single exponential equation. Data are represented as the mean ± SD from three independent replicates. Statistical significance was assessed using a two-tailed unpaired t-test.

Bioliography

[1] Stella, S.; Mesa, P.; Thomsen, J.; Paul, B.; Alcón, P.; Jensen, S. B.; Saligram, B.; Moses, M. E.; Hatzakis, N. S.; Montoya, G. Conformational Activation Promotes CRISPR-Cas12a Catalysis and Resetting of the Endonuclease Activity. *Cell* **2018**, *175*, 1856–1871.e21.

[2] Swarts, D. C.; van der Oost, J.; Jinek, M. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Molecular Cell* **2017**, *66*, 221–233.e4.

[3] Doudna, J. A.; Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **2014**, *346*, 1258096.

[4] Wright, A. V.; Nuñez, J. K.; Doudna, J. A. Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. *Cell* **2016**, *164*, 29–44.

[5] Barrangou, R.; Horvath, P. A decade of discovery: CRISPR functions and applications. *Nature Microbiology* **2017**, *2*, 1–9.

[6] Knipping, F.; Osborn, M. J.; Petri, K.; Tolar, J.; Glimm, H.; von Kalle, C.; Schmidt, M.; Gabriel, R. Genome-wide specificity of highly efficient TALENs and CRISPR/Cas9 for T cell receptor modification. *Molecular Therapy Methods & Clinical Development* **2017**, *4*, 213–224.

[7] Chandrasegaran, S.; Carroll, D. Origins of programmable nucleases for genome engineering. *Journal of Molecular Biology* **2016**, *428*, 963–989.

[8] Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; Charpentier, E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **2012**, *337*, 816–821.

[9] Gasiunas, G.; Barrangou, R.; Horvath, P.; Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences* **2012**, *109*, E2579–E2586.

[10] Cong, L.; Ran, F. A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A.; Zhang, F. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **2013**, *339*, 819–823.

[11] Wang, J. Y.; Pausch, P.; Doudna, J. A. Structural biology of CRISPR–Cas immunity and genome editing enzymes. *Nature Reviews Microbiology* **2022**,

[12] Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **2015**, *526*, 55–61.

[13] Jinek, M.; Jiang, F.; Taylor, D. W.; Sternberg, S. H.; Kaya, E.; Ma, E.; Anders, C.; Hauer, M.; Zhou, K.; Lin, S.; Kaplan, M.; Iavarone, A. T.; Charpentier, E.;

Nogales, E.; Doudna, J. A. Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science* **2014**, *343*, 1247997.

[14] Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* **2020**, *578*, 229–236.

[15] Zetsche, B.; Gootenberg, J. S.; Abudayyeh, O. O.; Slaymaker, I. M.; Makarova, K. S.; Essletzbichler, P.; Volz, S. E.; Joung, J.; van der Oost, J.; Regev, A.; Koonin, E. V.; Zhang, F. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **2015**, *163*, 759–771.

[16] Swarts, D. C.; Jinek, M. Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. *Molecular Cell* **2019**, *73*, 589–600.e4.

[17] Broughton, J. P. et al. CRISPR–Cas12-based detection of SARS-CoV-2. *Nature Biotechnology* **2020**, 1–5.

[18] Deng, F.; Li, Y.; Yang, B.; Sang, R.; Deng, W.; Kansara, M.; Lin, F.; Thavaneswaran, S.; Thomas, D. M.; Goldys, E. M. Topological barrier to Cas12a activation by circular DNA nanostructures facilitates autocatalysis and transforms DNA/RNA sensing. *Nature Communications* **2024**, *15*, 1818.

[19] Wani, A. K.; Akhtar, N.; Chopra, C.; Singh, R.; Hong, J. C.; Kadam, U. S.; others CRISPR/Cas12a-based biosensors for environmental monitoring and diagnostics. *Environmental Technology & Innovation* **2024**, 103625.

[20] Xu, X.; Dai, T.; Xiong, Q.; Yang, J.; Zang, J.; Liu, T. Molecular Detection of Phytophthora cinnamomi by RPA-CRISPR/Cas12a-Mediated Isothermal Amplification. *Forests* **2024**, *15*, 772.

[21] Nishimasu, H.; Nureki, O. Structures and mechanisms of CRISPR RNA-guided effector nucleases. *Current Opinion in Structural Biology* **2017**, *43*, 68–78.

[22] Yosef, I.; Goren, M. G.; Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research* **2012**, *40*, 5569–5576.

[23] Levy, A.; Goren, M. G.; Yosef, I.; Auster, O.; Manor, M.; Amitai, G.; Edgar, R.; Qimron, U.; Sorek, R. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **2015**, *520*, 505–510.

[24] Radovčić, M.; Killelea, T.; Savitskaya, E.; Wettstein, L.; Bolt, E. L.; Ivančić-Baće, I. CRISPR–Cas adaptation in Escherichia coli requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5' ssDNA exonucleases. *Nucleic Acids Research* **2018**, *46*, 10173–10183.

[25] Swarts, D. C.; Mosterd, C.; Van Passel, M. W.; Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. *PloS one* **2012**, *7*, e35888.

[26] Datsenko, K. A.; Pougach, K.; Tikhonov, A.; Wanner, B. L.; Severinov, K.; Semenova, E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications* **2012**, *3*, 945.

[27] Staals, R. H.; Jackson, S. A.; Biswas, A.; Brouns, S. J.; Brown, C. M.; Fineran, P. C. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. *Nature Communications* **2016**, *7*, 12853.

[28] Mojica, F. J.; Díez-Villaseñor, C. s.; García-Martínez, J.; Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution* **2005**, *60*, 174–182.

[29] Deltcheva, E.; Chylinski, K.; Sharma, C. M.; Gonzales, K.; Chao, Y.; Pirzada, Z. A.; Eckert, M. R.; Vogel, J.; Charpentier, E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **2011**, *471*, 602–607.

[30] Charpentier, E.; Richter, H.; van der Oost, J.; White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiology Reviews* **2015**, *39*, 428–441.

[31] Fonfara, I.; Richter, H.; BratoviÄ, M.; Rhun, A. L.; Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **2016**, *532*, 517–521.

[32] Westra, E. R.; Semenova, E.; Datsenko, K. A.; Jackson, R. N.; Wiedenheft, B.; Severinov, K.; Brouns, S. J. Type IE CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS genetics* **2013**, *9*, e1003742.

[33] Makarova, K. S.; Wolf, Y. I.; Alkhnbashi, O. S.; Costa, F.; Shah, S. A.; Saunders, S. J.; Barrangou, R.; Brouns, S. J.; Charpentier, E.; Haft, D. H.; others An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology* **2015**, *13*, 722–736.

[34] Brouns, S. J.; Jore, M. M.; Lundgren, M.; Westra, E. R.; Slijkhuis, R. J.; Snijders, A. P.; Dickman, M. J.; Makarova, K. S.; Koonin, E. V.; Van Der Oost, J. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **2008**, *321*, 960–964.

[35] Redding, S.; Sternberg, S. H.; Marshall, M.; Gibb, B.; Bhat, P.; Guegler, C. K.; Wiedenheft, B.; Doudna, J. A.; Greene, E. C. Surveillance and processing of foreign DNA by the Escherichia coli CRISPR-Cas system. *Cell* **2015**, *163*, 854–865.

[36] Swarts, D. C.; Jinek, M. Cas9 versus Cas12a/Cpf1: Structure-function comparisons and implications for genome editing. *Wiley Interdisciplinary Reviews: RNA* **2018**, *9*, e1481.

[37] Jiang, F.; Doudna, J. A. CRISPR–Cas9 Structures and Mechanisms. *Annual Review of Biophysics* **2017**, *46*, 505–529.

[38] Swartjes, T.; Staals, R. H. J.; van der Oost, J. Editor's cut: DNA cleavage by CRISPR RNA-guided nucleases Cas9 and Cas12a. *Biochemical Society Transactions* **2019**, *0*, 1–13.

[39] Shmakov, S.; Smargon, A.; Scott, D.; Cox, D.; Pyzocha, N.; Yan, W.; Abudayyeh, O. O.; Gootenberg, J. S.; Makarova, K. S.; Wolf, Y. I.; Severinov, K.; Zhang, F.; Koonin, E. V. Diversity and evolution of class 2 CRISPR–Cas systems. *Nature Reviews Microbiology* **2017**, *15*, 169–182.

[40] Stella, S.; Alcón, P.; Montoya, G. Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. *Nature* **2017**, *546*, 559–563.

[41] Yamano, T.; Nishimasu, H.; Zetsche, B.; Hirano, H.; Slaymaker, I.; Li, Y.; Fedorova, I.; Nakane, T.; Makarova, K.; Koonin, E.; Ishitani, R.; Zhang, F.; Nureki, O. Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **2016**, *165*, 949–962.

[42] Jeon, Y.; Choi, Y. H.; Jang, Y.; Yu, J.; Goo, J.; Lee, G.; Jeong, Y. K.; Lee, S. H.; Kim, I.-S.; Kim, J.-S.; Jeong, C.; Lee, S.; Bae, S. Direct observation of DNA target searching and cleavage by CRISPR-Cas12a. *Nature Communications* **2018**, *9*, 2777.

[43] Dong, D. et al. The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* **2016**, *532*, 522–526.

[44] Sternberg, S. H.; Redding, S.; Jinek, M.; Greene, E. C.; Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **2014**, *507*, 695a.

[45] Yamano, T.; Zetsche, B.; Ishitani, R.; Zhang, F.; Nishimasu, H.; Nureki, O. Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1. *Molecular Cell* **2017**, *67*, 633–645.

[46] Tu, M.; Lin, L.; Cheng, Y.; He, X.; Sun, H.; Xie, H.; Fu, J.; Liu, C.; Li, J.; Chen, D.; others A 'new lease of life': FnCpf1 possesses DNA cleavage activity for genome editing in human cells. *Nucleic Acids Research* **2017**, *45*, 11295–11304.

[47] Losito, M.; Smith, Q. M.; Newton, M. D.; Cuomo, M. E.; Rueda, D. S. Cas12a target search and cleavage on force-stretched DNA. *Physical Chemistry Chemical Physics* **2021**, *23*, 26640–26644.

[48] Strohkendl, I.; Saifuddin, F. A.; Rybarski, J. R.; Finkelstein, I. J.; Russell, R. Kinetic Basis for DNA Target Specificity of CRISPR-Cas12a. *Molecular Cell* **2018**, *71*, 816–824.e3.

[49] Stella, S.; Alcón, P.; Montoya, G. Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. *Nature* **2017**, *546*, 559–563.

[50] Anders, C.; Niewoehner, O.; Duerst, A.; Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **2014**, *513*, 569–573.

[51] Strohkendl, I.; Saha, A.; Moy, C.; Nguyen, A.-H.; Ahsan, M.; Russell, R.; Palermo, G.; Taylor, D. W. Cas12a domain flexibility guides R-loop formation and forces RuvC resetting. *Molecular Cell* **2024**,

[52] Wörle, E.; Jakob, L.; Schmidbauer, A.; Zinner, G.; Grohmann, D. Decoupling the bridge helix of Cas12a results in a reduced trimming activity, increased mismatch sensitivity and impaired conformational transitions. *Nucleic Acids Research* **2021**, *49*, 5278–5293.

[53] Naqvi, M. M.; Lee, L.; Montaguth, O. E. T.; Diffin, F. M.; Szczelkun, M. D. CRISPR–Cas12a-mediated DNA clamping triggers target-strand cleavage. *Nature Chemical Biology* **2022**, *18*, 1014–1022.

[54] Zhang, H.; Li, Z.; Daczkowski, C. M.; Gabel, C.; Mesecar, A. D.; Chang, L. Structural basis for the inhibition of CRISPR-Cas12a by anti-CRISPR proteins. *Cell host & Microbe* **2019**, *25*, 815–826.

[55] Singh, D.; Wang, Y.; Mallon, J.; Yang, O.; Fei, J.; Poddar, A.; Ceylan, D.; Bailey, S.; Ha, T. Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis. *Nature Structural & Molecular Biology* **2018**,

[56] Saha, A.; Ahsan, M.; Arantes, P. R.; Schmitz, M.; Chanez, C.; Jinek, M.; Palermo, G. An alpha-helical lid guides the target DNA toward catalysis in CRISPR-Cas12a. *Nature Communications* **2024**, *15*, 1473.

[57] Andreatta, D.; Sen, S.; Pérez Lustres, J. L.; Kovalenko, S. A.; Ernsting, N. P.; Murphy, C. J.; Coleman, R. S.; Berg, M. A. Ultrafast Dynamics in DNA:"Fraying" at the End of the Helix. *Journal of the American Chemical Society* **2006**, *128*, 6885–6892.

[58] Cofsky, J. C.; Karandur, D.; Huang, C. J.; Witte, I. P.; Kuriyan, J.; Doudna, J. A. CRISPR-Cas12a exploits R-loop asymmetry to form double-strand breaks. *Elife* **2020**, *9*, e55143.

[59] Chen, J. S.; Ma, E.; Harrington, L. B.; Costa, M. D.; Tian, X.; Palefsky, J. M.; Doudna, J. A. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **2018**, *360*, 436–439.

[60] Mohammad, N.; Talton, L.; Hetzler, Z.; Gongireddy, M.; Wei, Q. Unidirectional trans-cleaving behavior of CRISPR-Cas12a unlocks for an ultrasensitive assay using hybrid DNA reporters containing a 3′ toehold. *Nucleic Acids Research* **2023**, *51*, 9894–9904.

[61] Fuchs, R. T.; Curcuru, J.; Mabuchi, M.; Yourik, P.; Robb, G. B. Cas12a trans-cleavage can be modulated in vitro and is active on ssDNA, dsDNA, and RNA. *BioRxiv* **2019**, 600890.

[62] Zhang, L.; Sun, R.; Yang, M.; Peng, S.; Cheng, Y.; Chen, C. Conformational dynamics and cleavage sites of Cas12a are modulated by complementarity between crRNA and DNA. *iScience* **2019**, *19*, 492–503.

[63] Kim, H. K.; Song, M.; Lee, J.; Menon, A. V.; Jung, S.; Kang, Y.-M.; Choi, J. W.; Woo, E.; Koh, H. C.; Nam, J.-W.; others In vivo high-throughput profiling of CRISPR–Cpf1 activity. *Nature Methods* **2017**, *14*, 153–159.

[64] Fu, B. X. H.; Smith, J. D.; Fuchs, R. T.; Mabuchi, M.; Curcuru, J.; Robb, G. B.; Fire, A. Z. Target-dependent nickase activities of the CRISPR–Cas nucleases Cpf1 and Cas9. *Nature microbiology* **2019**, *4*, 888–897.

[65] Kleinstiver, B. P.; Tsai, S. Q.; Prew, M. S.; Nguyen, N. T.; Welch, M. M.; Lopez, J. M.; McCaw, Z. R.; Aryee, M. J.; Joung, J. K. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature Biotechnology* **2016**, *34*, 869–874.

[66] Kim, D.; Kim, J.; Hur, J. K.; Been, K. W.; Yoon, S.-h.; Kim, J.-S. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nature biotechnology* **2016**, *34*, 863–868.

[67] Chia, S.; Guo, T.; Goldys, E. M.; Payne, S. C.; Lovell, N. H.; Shivdasani, M. N.; Deng, F. A CRISPR mediated point-of-care assay for the detection of mucosal calprotectin in an animal model of ulcerative colitis. *medRxiv* **2024**, 2024–03.

[68] Li, S.-Y.; Cheng, Q.-X.; Liu, J.-K.; Nie, X.-Q.; Zhao, G.-P.; Wang, J. CRISPR-Cas12a has both cis-and trans-cleavage activities on single-stranded DNA. *Cell Research* **2018**, *28*, 491–493.

[69] Li, S.-Y.; Cheng, Q.-X.; Wang, J.-M.; Li, X.-Y.; Zhang, Z.-L.; Gao, S.; Cao, R.-B.; Zhao, G.-P.; Wang, J. CRISPR-Cas12a-assisted nucleic acid detection. *Cell discovery* **2018**, *4*, 20.

[70] Garcia-Doval, C.; Jinek, M. Molecular architectures and mechanisms of Class 2 CRISPR-associated nucleases. *Current Opinion in Structural Biology* **2017**, *47*, 157–166.

[71] Stella, S.; Alcón, P.; Montoya, G. Class 2 CRISPR–Cas RNA-guided endonucleases: Swiss Army knives of genome editing. *Nature Structural & Molecular Biology* **2017**, *24*, 882–892.

[72] Gao, P.; Yang, H.; Rajashankar, K. R.; Huang, Z.; Patel, D. J. Type v CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. *Cell Research* **2016**, *26*, 901–913.

[73] Palermo, G.; Miao, Y.; Walker, R. C.; Jinek, M.; McCammon, J. A. Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations. *ACS Central Science* **2016**, *2*, 756–763.

[74] Palermo, G.; Miao, Y.; Walker, R. C.; Jinek, M.; McCammon, J. A. CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proceedings of the National Academy of Sciences* **2017**, *114*, 7260–7265.

[75] Palermo, G.; Chen, J. S.; Ricci, C. G.; Rivalta, I.; Jinek, M.; Batista, V. S.; Doudna, J. A.; McCammon, J. A. Key role of the REC lobe during CRISPR–Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain. *Quarterly Reviews of Biophysics* **2018**, *51*, e9.

[76] Zuo, Z.; Zolekar, A.; Babu, K.; Lin, V. J. T.; Hayatshahi, H. S.; Rajan, R.; Wang, Y.-C.; Liu, J. Structural and functional insights into the bona fide catalytic state of Streptococcus pyogenes Cas9 HNH nuclease domain. *eLife* **2019**, *8*.

[77] Zuo, Z.; Weng, J.; Wang, W. Insights into the Inhibitory Mechanism of D13-9001 to the Multidrug Transporter AcrB through Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* **2016**, *120*, 2145–2154.

[78] Casalino, L.; Lukasz Nierzwicki; Jinek, M.; Palermo, G. Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR-Cas9 Revealed by Ab Initio Molecular Dynamics. *ACS Catalysis* **2020**, *10*, 13596–13605.

[79] Dagdas, Y. S.; Chen, J. S.; Sternberg, S. H.; Doudna, J. A. A Conformational Checkpoint between DNA Binding and Cleavage by CRISPR-Cas9. *Science Advances* **2017**, *3*, eaao002.

[80] Zhu, X.; Clarke, R.; Puppala, A. K.; Chittori, S.; Merk, A.; Merrill, B. J.; Simonović, M.; Subramaniam, S. Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nature Structural & Molecular Biology* **2019**, *26*, 679–685.

[81] Chen, J. S.; Dagdas, Y. S.; Kleinstiver, B. P.; Welch, M. M.; Sousa, A. A.; Harrington, L. B.; Sternberg, S. H.; Joung, J. K.; Yildiz, A.; Doudna, J. A. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **2017**, *550*, 407–410.

[82] Schmid-Burgk, J. L.; Gao, L.; Li, D.; Gardner, Z.; Strecker, J.; Lash, B.; Zhang, F. Highly parallel profiling of Cas9 variant specificity. *Molecular Cell* **2020**, *78*, 794–800.

[83] Yang, M.; Peng, S.; Sun, R.; Lin, J.; Wang, N.; Chen, C. The Conformational Dynamics of Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET. *Cell Reports* **2018**, *22*, 372–382.

[84] Casini, A.; Olivieri, M.; Petris, G.; Montagna, C.; Reginato, G.; Maule, G.; Lorenzin, F.; Prandi, D.; Romanel, A.; Demichelis, F.; Inga, A.; Cereseto, A. A highly specific spCas9 variant is identified by in vivo screening in yeast. *Nature Biotechnology* **2018**, *36*, 265–271.

[85] Kleinstiver, B. P.; Sousa, A. A.; Walton, R. T.; Tak, Y. E.; Hsu, J. Y.; Clement, K.; Welch, M. M.; Horng, J. E.; Malagon-Lopez, J.; Scarfò, I.; Maus, M. V.; Pinello, L.; Aryee, M. J.; Joung, J. K. Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nature Biotechnology* **2019**, *37*, 276–282.

[86] Yang, W.; Steitz, T. A. Recombining the structures of HIV integrase, RuvC and RNase H. *Structure* **1995**, *3*, 131–134.

[87] Daidone, I.; Amadei, A. Essential dynamics: foundation and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 762–770.

[88] Palermo, G.; Ricci, C. G.; Fernando, A.; Basak, R.; Jinek, M.; Rivalta, I.; Batista, V. S.; McCammon, J. A. Protospacer adjacent motif-induced allostery

activates CRISPR-Cas9. *Journal of the American Chemical Society* **2017**, *139*, 16028–16031.

[89] Jiang, F.; Doudna, J. A. The structural biology of CRISPR-Cas systems. *Current Opinion in Structural Biology* **2015**, *30*, 100–111.

[90] Gong, S.; Yu, H. H.; Johnson, K. A.; Taylor, D. W. DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell Reports* **2018**, *22*, 359–371.

[91] Liu, M. S.; Gong, S.; Yu, H. H.; Jung, K.; Johnson, K. A.; Taylor, D. W. Engineered CRISPR/Cas9 enzymes improve discrimination by slowing DNA cleavage to allow release of off-target DNA. *Nature Communications* **2020**, *11*, 1–13.

[92] Lange, O. F.; Grubmuller, H. Generalized correlation for biomolecular dynamics. *Proteins-Structure Function and Bioinformatics* **2006**, *62*, 1053–1061.

[93] Casalino, L.; Palermo, G.; Spinello, A.; Rothlisberger, U.; Magistrato, A. All-atom simulations disentangle the functional dynamics underlying gene maturation in the intron lariat spliceosome. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, 6584–6589.

[94] East, K. W.; Newton, J. C.; Morzan, U. N.; Narkhede, Y. B.; Acharya, A.; Skeens, E.; Jogl, G.; Batista, V. S.; Palermo, G.; Lisi, G. P. Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *Journal of the American Chemical Society* **2020**, *142*, 1348–1358.

[95] Lukasz Nierzwicki; Arantes, P. R.; Saha, A.; Palermo, G. Establishing the Allosteric Mecanism in CRISPR-Cas9. *WIREs Comput Mol Sci* **2020**, *In press*.

[96] Zuo, Z.; Liu, J. Allosteric regulation of CRISPR-Cas9 for DNA-targeting and cleavage. *Current Opinion in Structural Biology* **2020**, *62*, 166–174.

[97] Montefiori, M.; Pilotto, S.; Marabelli, C.; Moroni, E.; Ferraro, M.; Serapian, S. A.; Mattevi, A.; Colombo, G. Impact of Mutations on NPAC Structural Dynamics: Mechanistic Insights from MD Simulations. *Journal of Chemical Information and Modeling* **2019**, *59*, 3927–3937.

[98] Serapian, S. A.; Colombo, G. Designing Molecular Spanners to Throw in the Protein Networks. *Chemistry – A European Journal* **2020**, *26*, 4656–4670.

[99] Cui, Y.; Tang, Y.; Liang, M.; Ji, Q.; Zeng, Y.; Chen, H.; Lan, J.; Jin, P.; Wang, L.; Song, G.; Lou, J. Direct observation of the formation of a CRISPR–Cas12a R-loop complex at the single-molecule level. *Chemical Communications* **2020**, *56*, 2123–2126.

[100] Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; 3rd Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of Alpha/Gamma Conformers. *Biophysical Journal* **2007**, *92*, 3817–3829.

[101] Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; 3rd Cheatham, T. E.; Sponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for

RNA Simulations: Stability of UUCG and GNRA Hairpins. *Journal of Chemical Theory & Computation* **2010**, *6*, 3836–3849.

[102] Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E.; Jurecka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *Journal of Chemical Theory and Computation* **2011**, *7*, 2886–2902.

[103] Allnér, O.; Nilsson, L.; Villa, A. Magnesium ion–water coordination and exchange in biomolecular simulations. *Journal of Chemical Theory and Computation* **2012**, *8*, 1493–1502.

[104] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

[105] Ricci, C. G.; Chen, J. S.; Miao, Y.; Jinek, M.; Doudna, J. A.; McCammon, J. A.; Palermo, G. Deciphering Off-Target Effects in CRISPR-Cas9 through Accelerated Molecular Dynamics. *ACS Central Science* **2019**, *5*, 651–662.

[106] Palermo, G. Structure and dynamics of the CRISPR–Cas9 catalytic complex. *Journal of Chemical Information and Modeling* **2019**, *59*, 2394–2406.

[107] Turq, P.; Lantelme, F.; Friedman, H. L. Brownian Dynamics: Its Applications to Ionic Solutions. *J. Chem. Phys.* **1977**, *66*, 3039.

[108] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **1984**, *81*, 3684.

[109] Case, D. A. et al. AMBER 2018. *University of California, San Francisco* **2018**,

[110] Zhang, F. Development of CRISPR-Cas systems for genome editing and beyond. *Quarterly Reviews of Biophysics* **2019**, *52*, e6.

[111] Rossetti, M.; Merlo, R.; Bagheri, N.; Moscone, D.; Valenti, A.; Saha, A.; Arantes, P. R.; Ippodrino, R.; Ricci, F.; Treglia, I.; Delibato, E.; van der Oost, J.; Palermo, G.; Perugino, G.; Porchetta, A. Enhancement of CRISPR/-Cas12a ¡i¿trans¡/i¿ -cleavage activity using hairpin DNA reporters. *Nucleic Acids Research* **2022**,

[112] Paul, B.; Chaubet, L.; Verver, D. E.; Montoya, G. Mechanics of CRISPR-Cas12a and engineered variants on λ-DNA. *Nucleic Acids Research* **2022**, *50*, 5208–5225.

[113] Saha, A.; Arantes, P. R.; Hsu, R. V.; Narkhede, Y. B.; Jinek, M.; Palermo, G. Molecular Dynamics Reveals a DNA-Induced Dynamic Switch Triggering Activation of CRISPR-Cas12a. *Journal of Chemical Information and Modeling* **2020**, *60*, 6427–6437.

[114] Shaw, D. E. et al. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. 2014; pp 41–53.

[115] Bravo, J. P.; Liu, M.-S.; Hibshman, G. N.; Dangerfield, T. L.; Jung, K.; McCool, R. S.; Johnson, K. A.; Taylor, D. W. Structural basis for mismatch surveillance by CRISPR–Cas9. *Nature* **2022**, *603*, 343–347.

[116] Kumar, S.; Nussinov, R. Relationship between Ion Pair Geometries and Electrostatic Strengths in Proteins. *Biophysical Journal* **2002**, *83*, 1595–1612.

[117] Kästner, J. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 932–942.

[118] Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics* **2008**, *129*, 124105.

[119] Matsunaga, Y.; Kamiya, M.; Oshima, H.; Jung, J.; Ito, S.; Sugita, Y. Use of multistate Bennett acceptance ratio method for free-energy calculations from enhanced sampling and free-energy perturbation. *Biophysical Reviews* **2022**, *14*, 1503–1512.

[120] Pace, C. N.; Scholtz, J. M. A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophysical Journal* **1998**, *75*, 422–427.

[121] Lee, T.-S.; Lin, Z.; Allen, B. K.; Lin, C.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Sherman, W.; York, D. M. Improved alchemical free energy calculations with optimized smoothstep softcore potentials. *Journal of Chemical Theory and Computation* **2020**, *16*, 5512–5525.

[122] Bravo, J. P.; Hallmark, T.; Naegle, B.; Beisel, C. L.; Jackson, R. N.; Taylor, D. W. RNA targeting unleashes indiscriminate nuclease activity of CRISPR–Cas12a2. *Nature* **2023**, *613*, 582–587.

[123] Wörle, E.; Newman, A.; D'Silva, J.; Burgio, G.; Grohmann, D. Allosteric activation of CRISPR-Cas12a requires the concerted movement of the bridge helix and helix 1 of the RuvC II domain. *Nucleic Acids Research* **2022**, *50*, 10153–10168.

[124] Pausch, P.; Al-Shayeb, B.; Bisom-Rapp, E.; Tsuchida, C. A.; Li, Z.; Cress, B. F.; Knott, G. J.; Jacobsen, S. E.; Banfield, J. F.; Doudna, J. A. CRISPR-CasΦ from huge phages is a hypercompact genome editor. *Science* **2020**, *369*, 333–337.

[125] Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **2018**, *46*, W296–W303.

[126] Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation* **2020**, *16*, 528–552.

[127] Ivani, I. et al. Parmbsc1: a refined force field for DNA simulations. *Nature Methods* **2016**, *13*, 55–58.

[128] Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz Jr, K. M. Rational design of particle mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *Journal of Chemical Theory and Computation* **2013**, *9*, 2733–2748.

[129] Saha, A.; Arantes, P. R.; Palermo, G. Dynamics and mechanisms of CRISPR-Cas9 through the lens of computational methods. *Current Opinion in Structural Biology* **2022**, *75*, 102400.

[130] Nierzwicki, L.; East, K. W.; Binz, M. B.; Hsu, R. V.; Ahsan, M.; Arantes, P. R.; Skeens, E.; Pacesa, M.; Jinek, M.; Lisi, G. P.; Palermo, G. Principles of Target DNA Cleavage and Role of Mg2+ in the Catalysis of CRISPR-Cas9. *Nature Catalysis* **2022**, *accepted*.

[131] Case, D. A. et al. AMBER 2020. *University of California, San Francisco* **2020**,

[132] Tuckerman, M.; Berne, B. J.; Martyna, G. J. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics* **1992**, *97*, 1990–2001.

[133] Lippert, R. A.; Predescu, C.; Ierardi, D. J.; Mackenzie, K. M.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *The Journal of Chemical Physics* **2013**, *139*, 164106.

[134] Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **1994**, *101*, 4177–4189.

[135] Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **1985**, *31*, 1695–1697.

[136] Nosé, S. An extension of the canonical ensemble molecular dynamics method. *Molecular Physics* **1986**, *57*, 187–191.

[137] Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *The Journal of Chemical Physics* **2005**, *122*, 54101.

[138] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977**, *23*, 327–341.

[139] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021.

[140] Marcos-Alcalde, I.; Setoain, J.; Mendieta-Moreno, J. I.; Mendieta, J.; Gomez-Puertas, P. MEPSA: minimum energy pathway analysis for energy landscapes. *Bioinformatics* **2015**, *31*, 3853–3855.

[141] Hub, J. S.; De Groot, B. L.; Van Der Spoel, D. g_wham–A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *Journal of Chemical Theory and Computation* **2010**, *6*, 3713–3720.

[142] Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; Jr., T. D. M.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *Journal of Chemical Information and Modeling* **2020**, *60*, 5595–5623.

[143] Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 397–411.

[144] Wille, D.; Jaeger, H.; Schlaich, A.; Wu, Z.; Beckstein, O. Flamel. `https://github.com/alchemistry/flamel`, 2022.

[145] Mohanraju, P.; Oost, J.; Jinek, M.; Swarts, D. Heterologous Expression and Purification of the CRISPR-Cas12a/Cpf1 Protein. *Bio-protocol* **2018**, *8*.

[146] Kissling, L.; Monfort, A.; Swarts, D. C.; Wutz, A.; Jinek, M. *Preparation and electroporation of Cas12a/Cpf1-guide RNA complexes for introducing large gene deletions in mouse embryonic stem cells*; 2019; pp 241–263.

[147] Zhang, F.; Huang, Z. Mechanistic insights into the versatile class II CRISPR toolbox. *Trends in Biochemical Sciences* **2022**, *47*, 433–450.

[148] Knott, G. J.; Doudna, J. A. CRISPR-Cas guides the future of genetic engineering. 2018; `http://science.sciencemag.org/`.

[149] Raper, A. T.; Stephenson, A. A.; Suo, Z. Functional insights revealed by the kinetic mechanism of CRISPR/Cas9. *Journal of the American Chemical Society* **2018**, *140*, 2971–2984.

[150] Strohkendl, I.; Saifuddin, F. A.; Gibson, B. A.; Rosen, M. K.; Russell, R.; Finkelstein, I. J. Inhibition of CRISPR-Cas12a DNA targeting by nucleosomes and chromatin. *Science Advances* **2021**, *7*, eabd6030.

[151] Boyle, E. A.; Andreasson, J. O.; Chircus, L. M.; Sternberg, S. H.; Wu, M. J.; Guegler, C. K.; Doudna, J. A.; Greenleaf, W. J. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences* **2017**, *114*, 5461–5466.

[152] Zhang, L.; Rube, H. T.; Vakulskas, C. A.; Behlke, M. A.; Bussemaker, H. J.; Pufall, M. A. Systematic in vitro profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes. *Nucleic Acids Research* **2020**, *48*, 5037–5053.

[153] Jones, S. K.; Hawkins, J. A.; Johnson, N. V.; Jung, C.; Hu, K.; Rybarski, J. R.; Chen, J. S.; Doudna, J. A.; Press, W. H.; Finkelstein, I. J. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nature Biotechnology* **2021**, *39*, 84–93.

[154] Son, H.; Park, J.; Hwang, I.; Jung, Y.; Bae, S.; Lee, S. Mg2+-dependent conformational rearrangements of CRISPR-Cas12a R-loop complex are mandatory for complete double-stranded DNA cleavage. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2113747118.

[155] Adler, B. A.; Trinidad, M. I.; Bellieny-Rabelo, D.; Zhang, E.; Karp, H. M.; Skopintsev, P.; Thornton, B. W.; Weissman, R. F.; Yoon, P. H.; Chen, L.; others CasPEDIA Database: a functional classification system for class 2 CRISPR-Cas enzymes. *Nucleic Acids Research* **2024**, *52*, D590–D596.

[156] Koonin, E. V.; Gootenberg, J. S.; Abudayyeh, O. O. Discovery of diverse CRISPR-Cas systems and expansion of the genome engineering toolbox. *Biochemistry* **2023**, *62*, 3465–3487.

[157] Nakagawa, R.; Hirano, H.; Omura, S. N.; Nety, S.; Kannan, S.; Altae-Tran, H.; Yao, X.; Sakaguchi, Y.; Ohira, T.; Wu, W. Y.; others Cryo-EM structure of the transposon-associated TnpB enzyme. *Nature* **2023**, *616*, 390–397.

[158] Sasnauskas, G.; Tamulaitiene, G.; Druteika, G.; Carabias, A.; Silanskas, A.; Kazlauskas, D.; Venclovas, Č.; Montoya, G.; Karvelis, T.; Siksnys, V. TnpB structure reveals minimal functional core of Cas12 nuclease family. *Nature* **2023**, *616*, 384–389.

[159] Omura, S. N.; Nakagawa, R.; Südfeld, C.; Villegas Warren, R.; Wu, W. Y.; Hirano, H.; Laffeber, C.; Kusakizako, T.; Kise, Y.; Lebbink, J. H.; others Mechanistic and evolutionary insights into a type VM CRISPR–Cas effector enzyme. *Nature Structural & Molecular Biology* **2023**, *30*, 1172–1182.

[160] Nguyen, G. T.; Dhingra, Y.; Sashital, D. G. Miniature CRISPR-Cas12 endonucleases–Programmed DNA targeting in a smaller package. *Current Opinion in Structural Biology* **2022**, *77*, 102466.

[161] Künne, T.; Swarts, D. C.; Brouns, S. J. Planting the seed: target recognition of short guide RNAs. *Trends in Microbiology* **2014**, *22*, 74–83.

[162] Wiedenheft, B.; van Duijn, E.; Bultema, J. B.; Waghmare, S. P.; Zhou, K.; Barendregt, A.; Westphal, W.; Heck, A. J.; Boekema, E. J.; Dickman, M. J.; others RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences* **2011**, *108*, 10092–10097.

[163] Lewis, B. P.; Shih, I.-h.; Jones-Rhoades, M. W.; Bartel, D. P.; Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **2003**, *115*, 787–798.

[164] Sun, R.; Zhao, Y.; Wang, W.; Liu, J.-J. G.; Chen, C. Nonspecific interactions between Cas12a and dsDNA located downstream of the PAM mediate target search and assist AsCas12a for DNA cleavage. *Chemical Science* **2023**, *14*, 3839–3851.

[165] Phan, P. T.; Schelling, M.; Xue, C.; Sashital, D. G. *Methods in Enzymology*; Elsevier, 2019; Vol. 616; pp 61–85.

[166] Parameshwaran, H. P.; Babu, K.; Tran, C.; Guan, K.; Allen, A.; Kathiresan, V.; Qin, P. Z.; Rajan, R. The bridge helix of Cas12a imparts selectivity in cis-DNA cleavage and regulates trans-DNA cleavage. *FEBS Letters* **2021**, *595*, 892–912.

[167] Ma, E.; Chen, K.; Shi, H.; Stahl, E. C.; Adler, B.; Trinidad, M.; Liu, J.; Zhou, K.; Ye, J.; Doudna, J. A. Improved genome editing by an engineered CRISPR-Cas12a. *Nucleic Acids Research* **2022**, *50*, 12689–12701.

[168] Pacesa, M.; Lin, C.-H.; Cléry, A.; Saha, A.; Arantes, P. R.; Bargsten, K.; Irby, M. J.; Allain, F. H.-T.; Palermo, G.; Cameron, P.; others Structural basis for Cas9 off-target activity. *Cell* **2022**, *185*, 4067–4081.

[169] Huang, X.; Sun, W.; Cheng, Z.; Chen, M.; Li, X.; Wang, J.; Sheng, G.; Gong, W.; Wang, Y. Structural basis for two metal-ion catalysis of DNA cleavage by Cas12i2. *Nature Communications* **2020**, *11*, 5241.

[170] Zhang, H.; Li, Z.; Xiao, R.; Chang, L. Mechanisms for target recognition and cleavage by the Cas12i RNA-guided endonuclease. *Nature Structural & Molecular Biology* **2020**, *27*, 1069–1076.

[171] Zhang, B.; Luo, D.; Li, Y.; Perčulija, V.; Chen, J.; Lin, J.; Ye, Y.; Ouyang, S. Mechanistic insights into the R-loop formation and cleavage in CRISPR-Cas12i1. *Nature communications* **2021**, *12*, 3476.

[172] Yang, H.; Gao, P.; Rajashankar, K. R.; Patel, D. J. PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. *Cell* **2016**, *167*, 1814–1828.e12.

[173] Babin, V.; Roland, C.; Sagui, C. Adaptively biased molecular dynamics for free energy calculations. *The Journal of Chemical Physics* **2008**, *128*.

[174] Babin, V.; Karpusenka, V.; Moradi, M.; Roland, C.; Sagui, C. Adaptively biased molecular dynamics: An umbrella sampling method with a time-dependent potential. *International Journal of Quantum Chemistry* **2009**, *109*, 3666–3678.

[175] Sugimoto, N.; Nakano, S.-i.; Katoh, M.; Matsumura, A.; Nakamuta, H.; Ohmichi, T.; Yoneyama, M.; Sasaki, M. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **1995**, *34*, 11211–11216.

[176] Watkins Jr, N. E.; Kennelly, W. J.; Tsay, M. J.; Tuin, A.; Swenson, L.; Lee, H.-R.; Morosyuk, S.; Hicks, D. A.; SantaLucia Jr, J. Thermodynamic contributions of single internal rA· dA, rC· dC, rG· dG and rU· dT mismatches in RNA/DNA duplexes. *Nucleic Acids Research* **2011**, *39*, 1894–1902.

[177] Singh, J.; Liu, K. G.; Allen, A.; Jiang, W.; Qin, P. Z. A DNA unwinding equilibrium serves as a checkpoint for CRISPR-Cas12a target discrimination. *Nucleic Acids Research* **2023**, *51*, 8730–8743.

[178] Nguyen, G. T.; Schelling, M. A.; Buscher, K. A.; Sritharan, A.; Sashital, D. G. CRISPR-Cas12a exhibits metal-dependent specificity switching. *bioRxiv* **2023**, 2023–11.

[179] Tsuchida, C. A.; Zhang, S.; Doost, M. S.; Zhao, Y.; Wang, J.; O'Brien, E.; Fang, H.; Li, C.-P.; Li, D.; Hai, Z.-Y.; others Chimeric CRISPR-CasX enzymes and guide RNAs for improved genome editing activity. *Molecular Cell* **2022**, *82*, 1199–1209.

[180] Pacesa, M.; Loeff, L.; Querques, I.; Muckenfuss, L. M.; Sawicka, M.; Jinek, M. R-loop formation and conformational activation mechanisms of Cas9. *Nature* **2022**, *609*, 191–196.

[181] Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *Journal of structural biology* **2005**, *152*, 36–51.

[182] Zheng, S. Q.; Palovcak, E.; Armache, J.-P.; Verba, K. A.; Cheng, Y.; Agard, D. A. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods* **2017**, *14*, 331–332.

[183] Punjani, A.; Rubinstein, J. L.; Fleet, D. J.; Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **2017**, *14*, 290–296.

[184] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **2021**, *30*, 70–82.

[185] Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology* **2018**, *74*, 519–530.

[186] Emsley, P.; Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography* **2004**, *60*, 2126–2132.

[187] Adams, P. D. et al. PHENIX: A Comprehensive Python-based System for Macromolecular Structure Solution. *Acta Cryst. Sect. D.* **2010**, 213–221.

[188] Dillard, K. E.; Brown, M. W.; Johnson, N. V.; Xiao, Y.; Dolan, A.; Hernandez, E.; Dahlhauser, S. D.; Kim, Y.; Myler, L. R.; Anslyn, E. V.; others Assembly and

translocation of a CRISPR-Cas primed acquisition complex. *Cell* **2018**, *175*, 934–946.

[189] Minoukadeh, K.; Chipot, C.; Lelièvre, T. Potential of mean force calculations: a multiple-walker adaptive biasing force approach. *Journal of Chemical Theory and Computation* **2010**, *6*, 1008–1017.

[190] Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314*, 141–151.

[191] Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **2008**, *100*, 20603.

[192] Kaminski, M. M.; Abudayyeh, O. O.; Gootenberg, J. S.; Zhang, F.; Collins, J. J. CRISPR-based diagnostics. *Nature Biomedical Engineering* **2021**, *5*, 643–656.

[193] Abudayyeh, O. O.; Gootenberg, J. S.; Konermann, S.; Joung, J.; Slaymaker, I. M.; Cox, D. B. T.; Shmakov, S.; Makarova, K. S.; Semenova, E.; Minakhin, L.; Severinov, K.; Regev, A.; Lander, E. S.; Koonin, E. V.; Zhang, F. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **2016**, *353*.

[194] East-Seletsky, A.; O'Connell, M. R.; Knight, S. C.; Burstein, D.; Cate, J. H.; Tjian, R.; Doudna, J. A. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **2016**, *538*, 270–273.

[195] Yan, W. X.; Hunnewell, P.; Alfonse, L. E.; Carte, J. M.; Keston-Smith, E.; Sothiselvam, S.; Garrity, A. J.; Chong, S.; Makarova, K. S.; Koonin, E. V.; others Functionally diverse type V CRISPR-Cas systems. *Science* **2019**, *363*, 88–91.

[196] Tang, Y.; Gao, L.; Feng, W.; Guo, C.; Yang, Q.; Li, F.; Le, X. C. The CRISPR–Cas toolbox for analytical and diagnostic assay development. *Chemical Society Reviews* **2021**, *50*, 11844–11869.

[197] Kellner, M. J.; Koob, J. G.; Gootenberg, J. S.; Abudayyeh, O. O.; Zhang, F. SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nature Protocols* **2019**, *14*, 2986–3012.

[198] De Puig, H.; Lee, R. A.; Najjar, D.; Tan, X.; Soenksen, L. R.; Angenent-Mari, N. M.; Donghia, N. M.; Weckman, N. E.; Ory, A.; Ng, C. F.; others Minimally instrumented SHERLOCK (miSHERLOCK) for CRISPR-based point-of-care diagnosis of SARS-CoV-2 and emerging variants. *Science Advances* **2021**, *7*, eabh2944.

[199] Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **2017**, *356*, 438–442.

[200] Gootenberg, J. S.; Abudayyeh, O. O.; Kellner, M. J.; Joung, J.; Collins, J. J.; Zhang, F. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a and Csm6. *Science* **2018**, *360*, 439–444.

[201] Chertow, D. S. Next-generation diagnostics with CRISPR. *Science* **2018**, *360*, 381–382.

[202] Myhrvold, C.; Freije, C. A.; Gootenberg, J. S.; Abudayyeh, O. O.; Metsky, H. C.; Durbin, A. F.; Kellner, M. J.; Tan, A. L.; Paul, L. M.; Parham, L. A.; others Field-deployable viral diagnostics using CRISPR-Cas13. *Science* **2018**, *360*, 444–448.

[203] Barnes, K. G.; Lachenauer, A. E.; Nitido, A.; Siddiqui, S.; Gross, R.; Beitzel, B.; Siddle, K. J.; Freije, C. A.; Dighero-Kemp, B.; Mehta, S. B.; others Deployable CRISPR-Cas13a diagnostic tools to detect and report Ebola and Lassa virus cases in real-time. *Nature Communications* **2020**, *11*, 4131.

[204] Xiong, E.; Jiang, L.; Tian, T.; Hu, M.; Yue, H.; Huang, M.; Lin, W.; Jiang, Y.; Zhu, D.; Zhou, X. Simultaneous dual-gene diagnosis of SARS-CoV-2 based on CRISPR/Cas9-mediated lateral flow assay. *Angewandte Chemie International Edition* **2021**, *60*, 5307–5315.

[205] Mukama, O.; Wu, J.; Li, Z.; Liang, Q.; Yi, Z.; Lu, X.; Liu, Y.; Liu, Y.; Hussain, M.; Makafe, G. G.; others An ultrasensitive and specific point-of-care CRISPR/Cas12 based lateral flow biosensor for the rapid detection of nucleic acids. *Biosensors and Bioelectronics* **2020**, *159*, 112143.

[206] Wang, X.; Ji, P.; Fan, H.; Dang, L.; Wan, W.; Liu, S.; Li, Y.; Yu, W.; Li, X.; Ma, X.; others CRISPR/Cas12a technology combined with immunochromatographic strips for portable detection of African swine fever virus. *Communications Biology* **2020**, *3*, 62.

[207] Ramachandran, A.; Santiago, J. G. CRISPR enzyme kinetics for molecular diagnostics. *Analytical Chemistry* **2021**, *93*, 7456–7464.

[208] Choi, J.-H.; Lim, J.; Shin, M.; Paek, S.-H.; Choi, J.-W. CRISPR-Cas12a-based nucleic acid amplification-free DNA biosensor via Au nanoparticle-assisted metal-enhanced fluorescence and colorimetric analysis. *Nano Letters* **2021**, *21*, 693–699.

[209] Fozouni, P. et al. Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. *Cell* **2021**, *184*, 323–333.e9.

[210] Tian, T.; Shu, B.; Jiang, Y.; Ye, M.; Liu, L.; Guo, Z.; Han, Z.; Wang, Z.; Zhou, X. An ultralocalized Cas13a assay enables universal and nucleic acid amplification-free single-molecule RNA diagnostics. *ACS Nano* **2021**, *15*, 1167–1178.

[211] Bruch, R.; Johnston, M.; Kling, A.; Mattmüller, T.; Baaske, J.; Partel, S.; Madlener, S.; Weber, W.; Urban, G. A.; Dincer, C. CRISPR-powered electrochemical microfluidic multiplexed biosensor for target amplification-free miRNA diagnostics. *Biosensors and Bioelectronics* **2021**, *177*, 112887.

[212] Bao, M.; Chen, Q.; Xu, Z.; Jensen, E. C.; Liu, C.; Waitkus, J. T.; Yuan, X.; He, Q.; Qin, P.; Du, K. Challenges and opportunities for clustered regularly interspaced short palindromic repeats based molecular biosensing. *ACS Sensors* **2021**, *6*, 2497–2522.

[213] Bruch, R.; Baaske, J.; Chatelle, C.; Meirich, M.; Madlener, S.; Weber, W.; Dincer, C.; Urban, G. A. CRISPR/Cas13a-powered electrochemical microfluidic biosensor for nucleic acid amplification-free miRNA diagnostics. *Advanced Materials* **2019**, *31*, 1905311.

[214] Lillis, L.; Siverson, J.; Lee, A.; Cantera, J.; Parker, M.; Piepenburg, O.; Lehman, D. A.; Boyle, D. S. Factors influencing recombinase polymerase amplification (RPA) assay outcomes at point of care. *Molecular and Cellular Probes* **2016**, *30*, 74–78.

[215] Li, L.; Li, S.; Wu, N.; Wu, J.; Wang, G.; Zhao, G.; Wang, J. HOLMESv2: a CRISPR-Cas12b-assisted platform for nucleic acid detection and DNA methylation quantitation. *ACS Synthetic Biology* **2019**, *8*, 2228–2237.

[216] Steens, J. A.; Zhu, Y.; Taylor, D. W.; Bravo, J. P.; Prinsen, S. H.; Schoen, C. D.; Keijser, B. J.; Ossendrijver, M.; Hofstra, L. M.; Brouns, S. J.; others SCOPE enables type III CRISPR-Cas diagnostics using flexible targeting and stringent CARF ribonuclease activation. *Nature Communications* **2021**, *12*, 5033.

[217] Santiago-Frangos, A.; Hall, L. N.; Nemudraia, A.; Nemudryi, A.; Krishna, P.; Wiegand, T.; Wilkinson, R. A.; Snyder, D. T.; Hedges, J. F.; Cicha, C.; others Intrinsic signal amplification by type III CRISPR-Cas systems provides a sequence-specific SARS-CoV-2 diagnostic. *Cell Reports Medicine* **2021**, *2*.

[218] Liu, T. Y.; Knott, G. J.; Smock, D. C.; Desmarais, J. J.; Son, S.; Bhuiya, A.; Jakhanwal, S.; Prywes, N.; Agrawal, S.; Díaz de León Derby, M.; others Accelerated RNA detection using tandem CRISPR nucleases. *Nature Chemical Biology* **2021**, *17*, 982–988.

[219] Nguyen, L. T.; Smith, B. M.; Jain, P. K. Enhancement of trans-cleavage activity of Cas12a with engineered crRNA enables amplified nucleic acid detection. *Nature Communications* **2020**, *11*, 4906.

[220] Nalefski, E. A.; Patel, N.; Leung, P. J.; Islam, Z.; Kooistra, R. M.; Parikh, I.; Marion, E.; Knott, G. J.; Doudna, J. A.; Le Ny, A.-L. M.; others Kinetic analysis of Cas12a and Cas13a RNA-Guided nucleases for development of improved CRISPR-Based diagnostics. *iScience* **2021**, *24*.

[221] Fu, X.; Shi, Y.; Peng, F.; Zhou, M.; Yin, Y.; Tan, Y.; Chen, M.; Yin, X.; Ke, G.; Zhang, X.-B. Exploring the trans-cleavage activity of CRISPR/Cas12a on gold nanoparticles for stable and sensitive biosensing. *Analytical Chemistry* **2021**, *93*, 4967–4974.

[222] Dai, Y.; Somoza, R. A.; Wang, L.; Welter, J. F.; Li, Y.; Caplan, A. I.; Liu, C. C. Exploring the trans-cleavage activity of CRISPR-Cas12a (cpf1) for the development of a universal electrochemical biosensor. *Angewandte Chemie* **2019**, *131*, 17560–17566.

[223] Li, Y.; Li, T.; Liu, B.-F.; Hu, R.; Zhu, J.; He, T.; Zhou, X.; Li, C.; Yang, Y.; Liu, M. CRISPR-Cas12a trans-cleaves DNA G-quadruplexes. *Chemical Communications* **2020**, *56*, 12526–12529.

[224] Smith, C. W.; Nandu, N.; Kachwala, M. J.; Chen, Y.-S.; Uyar, T. B.; Yigit, M. V. Probing CRISPR-Cas12a nuclease activity using double-stranded DNA-templated fluorescent substrates. *Biochemistry* **2020**, *59*, 1474–1481.

[225] Zhang, D.; Yan, Y.; Que, H.; Yang, T.; Cheng, X.; Ding, S.; Zhang, X.; Cheng, W. CRISPR/Cas12a-mediated interfacial cleaving of hairpin DNA reporter for electrochemical nucleic acid sensing. *ACS Sensors* **2020**, *5*, 557–562.

[226] Marqués, M.-C.; Ruiz, R.; Montagud-Martínez, R.; Márquez-Costa, R.; Albert, S.; Domingo-Calap, P.; Rodrigo, G. CRISPR-Cas12a-based detection of SARS-CoV-2 harboring the E484K mutation. *ACS Synthetic Biology* **2021**, *10*, 3595–3599.

[227] Mergny, J.-L.; Lacroix, L. Analysis of thermal melting curves. *Oligonucleotides* **2003**, *13*, 515–537.

[228] Idili, A.; Ricci, F.; Vallée-Bélisle, A. Determining the folding and binding free energy of DNA-based nanodevices and nanoswitches using urea titration curves. *Nucleic Acids Research* **2017**, *45*, 7571–7580.

[229] Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *Journal of Chemical Theory and Computation* **2007**, *3*, 156–169.

[230] Van Zundert, G.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastritis, P.; Karaca, E.; Melquiond, A.; van Dijk, M.; De Vries, S.; Bonvin, A. The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of Molecular Biology* **2016**, *428*, 720–725.

[231] Braselmann, E.; Rathbun, C.; Richards, E. M.; Palmer, A. E. Illuminating RNA biology: tools for imaging RNA in live mammalian cells. *Cell Chemical Biology* **2020**, *27*, 891–903.

[232] Tyagi, S. Imaging intracellular RNA distribution and dynamics in living cells. *Nature Methods* **2009**, *6*, 331–338.

[233] Kim, Y.; Sohn, D.; Tan, W. Molecular beacons in biomedical detection and clinical diagnosis. *International Journal of Clinical and Experimental Pathology* **2008**, *1*, 105.

[234] Zanut, A.; Rossetti, M.; Marcaccio, M.; Ricci, F.; Paolucci, F.; Porchetta, A.; Valenti, G. DNA-Based nanoswitches: insights into electrochemiluminescence signal enhancement. *Analytical Chemistry* **2021**, *93*, 10397–10402.

[235] Rossetti, M.; Ippodrino, R.; Marini, B.; Palleschi, G.; Porchetta, A. Antibody-mediated small molecule detection using programmable DNA-switches. *Analytical Chemistry* **2018**, *90*, 8196–8201.

[236] Porchetta, A.; Ippodrino, R.; Marini, B.; Caruso, A.; Caccuri, F.; Ricci, F. Programmable nucleic acid nanoswitches for the rapid, single-step detection of antibodies in bodily fluids. *Journal of the American Chemical Society* **2018**, *140*, 947–953.

[237] Mao, S.; Ying, Y.; Wu, R.; Chen, A. K. Recent advances in the molecular beacon technology for live-cell single-molecule imaging. *iScience* **2020**, *23*.

[238] Vallée-Bélisle, A.; Ricci, F.; Plaxco, K. W. Thermodynamic basis for the optimization of binding-induced biomolecular switches and structure-switching biosensors. *Proceedings of the National Academy of Sciences* **2009**, *106*, 13802–13807.

[239] Farag, N.; Mattossovich, R.; Merlo, R.; Lukasz Nierzwicki; Palermo, G.; Porchetta, A.; Perugino, G.; Ricci, F. Folding-upon-Repair DNA Nanoswitches

for Monitoring the Activity of DNA Repair Enzymes. *Angewandte Chemie* **2021**, *133*, 7359–7365.

[240] Mocenigo, M.; Porchetta, A.; Rossetti, M.; Brass, E.; Tonini, L.; Puzzi, L.; Tagliabue, E.; Triulzi, T.; Marini, B.; Ricci, F.; others Rapid, cost-effective peptide/nucleic acid-based platform for therapeutic antibody monitoring in clinical samples. *ACS Sensors* **2020**, *5*, 3109–3115.

[241] Mariottini, D.; Idili, A.; Nijenhuis, M. A.; de Greef, T. F.; Ricci, F. DNA-based nanodevices controlled by purely entropic linker domains. *Journal of the American Chemical Society* **2018**, *140*, 14725–14734.

[242] Mariottini, D.; Idili, A.; Nijenhuis, M. A.; Ercolani, G.; Ricci, F. Entropy-based rational modulation of the pKa of a synthetic pH-dependent nanoswitch. *Journal of the American Chemical Society* **2019**, *141*, 11367–11371.

[243] Rodrigues, J. P.; Bonvin, A. M. Integrative computational modeling of protein interactions. *The FEBS Journal* **2014**, *281*, 1988–2003.

[244] Hacisuleyman, E.; Hale, C.; Saito, Y.; Blachere, N. E.; Bergh, M.; Conlon, E. G.; Schaefer-Babajew, D. J.; DaSilva, J.; Muecksch, F.; Gaebler, C.; others Vaccine breakthrough infections with SARS-CoV-2 variants. *New England Journal of Medicine* **2021**, *384*, 2212–2218.

[245] Bell, R. L.; Jarvis, K. G.; Ottesen, A. R.; McFarland, M. A.; Brown, E. W. Recent and emerging innovations in Salmonella detection: a food and environmental perspective. *Microbial Biotechnology* **2016**, *9*, 279–292.

[246] Rodriguez-Lazaro, D.; Gonzalez-García, P.; Delibato, E.; De Medici, D.; García-Gimeno, R. M.; Valero, A.; Hernandez, M. Next day Salmonella spp. detection method based on real-time PCR for meat, dairy and vegetable food products. *International Journal of Food Microbiology* **2014**, *184*, 113–120.

[247] Malorny, B.; Paccassoni, E.; Fach, P.; Bunge, C.; Martin, A.; Helmuth, R. Diagnostic real-time PCR for detection of Salmonella in food. *Applied and Environmental Microbiology* **2004**, *70*, 7046–7052.

[248] Wang, J. Y.; Doudna, J. A. CRISPR technology: A decade of genome editing is only the beginning. *Science* **2023**, *379*, eadd8643.

[249] Mali, P.; Yang, L.; Esvelt, K. M.; Aach, J.; Guell, M.; DiCarlo, J. E.; Norville, J. E.; Church, G. M. RNA-Guided Human Genome Engineering via Cas9. *Science* **2013**, *339*, 823–826.

[250] Xue, C.; Greene, E. C. DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing. *Trends in Genetics* **2021**, *37*, 639–656.

[251] Anzalone, A. V.; Koblan, L. W.; Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. 2020.

[252] Liao, J.; Chen, S.; Hsiao, S.; Jiang, Y.; Yang, Y.; Zhang, Y.; Wang, X.; Lai, Y.; Bauer, D. E.; Wu, Y. Therapeutic adenine base editing of human hematopoietic stem cells. *Nature Communications* **2023**, *14*, 207.

[253] 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* **2016**, *533*, 420–424.

[254] Gaudelli, N. M.; Komor, A. C.; Rees, H. A.; Packer, M. S.; Badran, A. H.; Bryson, D. I.; Liu, D. R. Programmable base editing of T to G C in genomic DNA without DNA cleavage. *Nature* **2017**, *551*, 464–471.

[255] Levy, J. M.; Yeh, W.-H.; Pendse, N.; Davis, J. R.; Hennessey, E.; Butcher, R.; Koblan, L. W.; Comander, J.; Liu, Q.; Liu, D. R. Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nature Biomedical Engineering* **2020**, *4*, 97–110.

[256] Lam, D. K. et al. Improved cytosine base editors generated from TadA variants. *Nature Biotechnology* **2023**, *41*, 686–697.

[257] Neugebauer, M. E.; Hsu, A.; Arbab, M.; Krasnow, N. A.; McElroy, A. N.; Pandey, S.; Doman, J. L.; Huang, T. P.; Raguram, A.; Banskota, S.; others Evolution of an adenine base editor into a small, efficient cytosine base editor with low off-target activity. *Nature Biotechnology* **2023**, *41*, 673–685.

[258] Chen, X.; McAndrew, M. J.; Lapinaite, A. Unlocking the secrets of ABEs: the molecular mechanism behind their specificity. *Biochemical Society Transactions* **2023**, *51*, 1635–1646.

[259] Lapinaite, A.; Knott, G. J.; Palumbo, C. M.; Lin-Shiao, E.; Richter, M. F.; Zhao, K. T.; Beal, P. A.; Liu, D. R.; Doudna, J. A. DNA capture by a CRISPR-Cas9–guided adenine base editor. *Science* **2020**, *369*, 566–571.

[260] Wolf, J. tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *The EMBO Journal* **2002**, *21*, 3841–3851.

[261] Kim, J.; Malashkevich, V.; Roday, S.; Lisbin, M.; Schramm, V. L.; Almo, S. C. Structural and Kinetic Characterization of ¡i¿Escherichia coli¡/i¿ TadA, the Wobble-Specific tRNA Deaminase. *Biochemistry* **2006**, *45*, 6407–6416.

[262] Richter, M. F.; Zhao, K. T.; Eton, E.; Lapinaite, A.; Newby, G. A.; Thuronyi, B. W.; Wilson, C.; Koblan, L. W.; Zeng, J.; Bauer, D. E.; Doudna, J. A.; Liu, D. R. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nature Biotechnology* **2020**, *38*, 883–891.

[263] Xiao, Y.-L.; Wu, Y.; Tang, W. An adenine base editor variant expands context compatibility. *Nature Biotechnology* **2024**, 1–12.

[264] Losey, H. C.; Ruthenburg, A. J.; Verdine, G. L. Crystal structure of Staphylococcus aureus tRNA adenosine deaminase TadA in complex with RNA. *Nature Structural and Molecular Biology* **2006**, *13*, 153–159.

[265] Nguyen, H.; Pérez, A.; Bermeo, S.; Simmerling, C. Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins. *Journal of Chemical Theory and Computation* **2015**, *11*, 3714–3728.

[266] Nguyen, H.; Roe, D. R.; Simmerling, C. Improved Generalized Born Solvent Model Parameters for Protein Simulations. *Journal of Chemical Theory and Computation* **2013**, *9*, 2020–2034.

[267] Knight, J. L.; Brooks III, C. L. λ-Dynamics free energy simulation methods. *Journal of Computational Chemistry* **2009**, *30*, 1692–1700.

[268] Seeliger, D.; de Groot, B. L. Protein thermostability calculations using alchemical free energy simulations. *Biophysical journal* **2010**, *98*, 2309–2316.

[269] Pitera, J. W.; Kollman, P. A. Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. *Proteins: Structure, Function, and Bioinformatics* **2000**, *41*, 385–397.

[270] Romero, P. A.; Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **2009**, *10*, 866–876.

[271] Tokuriki, N.; Stricher, F.; Serrano, L.; Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Computational Biology* **2008**, *4*, e1000002.

[272] Wang, X.; Minasov, G.; Shoichet, B. K. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. *Journal of Molecular Biology* **2002**, *320*, 85–95.

[273] Rallapalli, K. L.; Ranzau, B. L.; Ganapathy, K. R.; Komor, A. C.; Paesani, F. Retracing the evolutionary trajectory of adenine base editors using theoretical approaches. *bioRxiv* **2020**, 2020.12.24.424366.

[274] Otten, R.; Liu, L.; Kenner, L. R.; Clarkson, M. W.; Mavor, D.; Tawfik, D. S.; Kern, D.; Fraser, J. S. Rescue of conformational dynamics in enzyme catalysis by directed evolution. *Nature Communications* **2018**, *9*, 1314.

[275] Corbella, M.; Pinto, G. P.; Kamerlin, S. C. Loop dynamics and the evolution of enzyme activity. *Nature Reviews Chemistry* **2023**, *7*, 536–547.

[276] Zhu, H.; Wang, L.; Wang, Y.; Jiang, X.; Qin, Q.; Song, M.; Huang, Q. Directed-evolution mutations enhance DNA-binding affinity and protein stability of the adenine base editor ABE8e. *Cellular and Molecular Life Sciences* **2024**, *81*, 257.

[277] Ranzau, B. L.; Rallapalli, K. L.; Evanoff, M.; Paesani, F.; Komor, A. C. The Wild-Type tRNA Adenosine Deaminase Enzyme TadA Is Capable of Sequence-Specific DNA Base Editing. *ChemBioChem* **2023**, *24*.

[278] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[279] Sinha, S.; Pindi, C.; Ahsan, M.; Arantes, P. R.; Palermo, G. Machines on Genes through the Computational Microscope. *Journal of Chemical Theory and Computation* **2023**, *19*, 1945–1964.

[280] Sinha, S.; Molina Vargas, A.; Arantes, P.; Patel, A.; O'Connell, M.; Palermo, G. Unveiling the RNA-mediated allosteric activation discloses functional hotspots in CRISPR-Cas13a. *Nucleic Acids Research* **2024**, *52*, 906–920.

[281] Galindo-Murillo, R.; Robertson, J. C.; Zgarbova, M.; Sponer, J.; Otyepka, M.; Jurecka, P.; Cheatham III, T. E. Assessing the current state of amber force field

modifications for DNA. *Journal of chemical theory and computation* **2016**, *12*, 4114–4127.

[282] Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *The Journal of Chemical Physics* **1993**, *98*, 1372–1377.

[283] Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *The Journal of Chemical Physics* **1984**, *80*, 3265–3269.

[284] Li, P.; Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *Journal of Chemical Information and Modeling* **2016**, *56*, 599–604.

[285] Li, P.; Merz, K. M. *Parameterization of a Dioxygen Binding Metal Site Using the MCPB.py Program*; 2021; pp 257–275.

[286] Lee, T.-S.; Radak, B. K.; Pabis, A.; York, D. M. A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *Journal of Chemical Theory and Computation* **2013**, *9*, 153–164.

[287] Wieczór, M.; Czub, J. Telomere uncapping by common oxidative guanine lesions: Insights from atomistic models. *Free Radical Biology and Medicine* **2020**, *148*, 162–169.

[288] Gapsys, V.; de Groot, B. L. pmx Webserver: A User Friendly Interface for Alchemistry. *Journal of Chemical Information and Modeling* **2017**, *57*, 109–114.

[289] Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L.; Schwede, T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research* **2014**, *42*, W252–W258.

[290] Palermo, G.; Cavalli, A.; Klein, M. L.; Alfonso-Prieto, M.; Peraro, M. D.; Vivo, M. D. Catalytic Metal Ions and Enzymatic Processing of DNA and RNA. *Accounts of Chemical Research* **2015**, *48*, 220–228.

[291] Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics* **1993**, *17*, 412–425.

[292] Ricci, C. G.; Silveira, R. L.; Rivalta, I.; Batista, V. S.; Skaf, M. S. Allosteric Pathways in the PPARγ-RXRα nuclear receptor complex. *Scientific Reports* **2016**, *6*, 19940.

[293] Bakan, A.; Meireles, L. M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27*, 1575–1577.

[294] Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **1996**, *14*, 27–28, 33–38.

[295] Wagner, J. R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R. E. POVME 3.0: Software for Mapping Binding Pocket Flexibility. *Journal of Chemical Theory and Computation* **2017**, *13*, 4584–4592.

[296] Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **2005**, *26*, 1701–1718.

[297] Floyd, R. W. Algorithm 97: Shortest path. *Communications of the ACM* **1962**, *5*, 345.

[298] Fiser, A.; Šali, A. *Modeller: Generation and Refinement of Homology-Based Protein Structure Models*; 2003; pp 461–491.

[299] Yang, W. Nucleases: diversity of structure, function and mechanism. *Quarterly Reviews of Biophysics* **2011**, *44*, 1–93.

[300] Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A.; Krieger, E.; Joosten, R. P.; Vriend, G. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research* **2015**, *43*, D364–D368.

[301] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.

[302] Punjani, A. Real-time cryo-EM structure determination. *Microscopy and Microanalysis* **2021**, *27*, 1156–1157.

[303] Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. ¡i¿MMPBSA.py¡/i¿ : An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation* **2012**, *8*, 3314–3321.

[304] Case, D. A. et al. AMBER 2022. *University of California, San Francisco* **2022**,

[305] Jiang, F.; Taylor, D. W.; Chen, J. S.; Kornfeld, J. E.; Zhou, K.; Thompson, A. J.; Nogales, E.; Doudna, J. A. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **2016**, *351*, 867–871.