Unveiling the Structural Determinants for Enhanced Specificity in CRISPR/Cas9 for Genome Editing

By:

Gayatri Panda

PhD19206

Under the supervision of Dr. Arjun Ray

Department of Computational Biology

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

August, 2025

Unveiling the Structural Determinants for Enhanced Specificity in CRISPR/Cas9 for Genome Editing

A Thesis

Submitted in Partial Fulfilment of the Requirements for the Degree Of

Doctor of Philosophy

By

Gayatri Panda

PhD19206

Under the supervision of Dr. Arjun Ray

Department of Computational Biology

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

August, 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Unveiling the Structural Determinants for Enhanced Specificity in CRISPR/Cas9 for Genome Editing submitted to IIIT-Delhi**, submitted by **Gayatri Panda (PhD19206)**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by her under the supervision of **Dr. Arjun Ray**. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Arjun Ray
Assistant Professor
Dept. of Computational Biology
IIIT Delhi, 110020
Place: New Delhi
Date: 29th August 2025

ACKNOWLEDGEMENTS

"Perseverance is not a long race; it is many short races one after the other." - Walter Elliot

A thesis is not just the culmination of years of work, but the sum of countless lessons, people, and moments that shape who we become. This stage represents a significant milestone in my life as I approach the completion of my doctoral thesis. It is a moment filled with profound emotions, satisfaction at bringing an important phase of my academic journey to a close, and a sense of nostalgia for the people and experiences that have profoundly influenced my career and personal growth. My years at the Indraprastha Institute of Information Technology, Delhi (IIIT-D) have been transformative, shaping both my career and my outlook on life. I am deeply grateful to everyone who has supported me along the way. A doctoral thesis is not merely an academic exercise; it is a test of perseverance, patience, and resilience. Above all, it requires the wisdom, guidance, and encouragement of a mentor. In my case, I have been fortunate to have **Dr. Arjun Ray** as my advisor. His vision, intellect, and commitment have been a constant source of inspiration. He has always encouraged me to think beyond the obvious, welcomed my ideas with openness, and provided an environment of academic freedom and rigor. His constant support, whether during moments of success or struggle, has been invaluable. His belief in my potential, especially during times when I felt low and lost confidence, has been a source of immense strength for me. I remain deeply indebted to him for giving me the opportunity to be part of his research team for these past five years, and for shaping not only my research but also my outlook on learning and life.

I would like to express my heartfelt gratitude to my parents, **Mr. Vasudeva Panda** and **Mrs. Rashmi Kumari Panda**, and my elder brother, **Mr. Prabhakar Panda**, for their unwavering love, support, and encouragement throughout my academic journey. My parents have been my pillars of strength, always providing a calming presence during stressful times and reminding me of the importance of resilience and hard work. Their sacrifices and dedication to ensuring my success have motivated me to persevere and strive for excellence. To my elder brother, Prabhakar, I extend my deepest thanks for being a constant source of encouragement and inspiration. His belief in

my potential and his reassuring words during moments of self-doubt have given me the courage to overcome challenges and stay focused on my goals. This journey would not have been possible without the unconditional love, understanding, and encouragement of my family.

I would also like to express special thanks to my collaborators who played a crucial role in my research endeavors. My sincere appreciation goes to **Prof. Vinod Scaria** for the opportunity to work on my first collaborative project at IIITD focused on SARS-CoV-2. I am grateful to **Prof. Mitali** for allowing me to collaborate on her project, which enriched my research experience. A special mention goes to **Prof. Satarupa Banerjee**, whose collaboration provided me with the opportunity for first authorship alongside her talented students. I am indebted to **Prof. Debojyoti Chakraborty** for his constant guidance throughout my CRISPR-related projects and throughout my entire thesis journey.

I would also like to extend my gratitude to my examiners and faculty members, whose feedback and encouragement have significantly enriched my research. Special thanks to **Dr. Jaspreet Kaur**, who not only guided and encouraged me but also shared valuable insights from her own PhD journey in CRISPR, which greatly benefited my work. I am grateful to **Prof. Ganesh Bagler**, **Dr. Tarini Shankar Ghosh**, and **Prof. Vibhor Kumar** for their invaluable suggestions and constructive feedback, which have helped me improve the quality of my work. Thanks to **Prof. Debarka Sengupta** and **Prof. Arul Murgan**, whose technical guidance and thoughtful suggestions have been pivotal in shaping my research. Your support and expertise have been truly instrumental in helping me navigate this journey.

I would like to acknowledge the unwavering support of my fellow students, starting with **Neha Mishra**, who holds a special place as not only my first friend and co-author but also as someone who guided me through the initial and often challenging phase of my PhD journey. Despite being an MTech student, Neha was always there to lend her wisdom and support, making initial years of my PhD both productive and comforting. Her camaraderie brought warmth and positivity to the lab, making it a welcoming space that I fondly remember. I want to express my gratitude to **Sukriti Sacher,** with whom I have spent a significant part of my PhD journey, four memorable years. While she is six months junior to me, we have shared a collegial bond that has grown stronger over time. Sukriti often took thoughtful initiatives in our lab, contributing meaningfully to its environment. I fondly remember the moments we shared, from engaging discussions to the enjoyable evenings spent at her place, where her hospitality made those times truly special. I deeply appreciate the friendship and memories we have created together during this journey. I would like to express my heartfelt gratitude to **Himanshi Aggarwal** and **Shivansh Verma**, the

MTech graduates from my lab, who have been a great support in my projects. Himanshi, in particular, has been an integral part of my journey, and I am grateful for the good bond we continue to share.

I would like to express my heartfelt gratitude to my junior lab colleagues **Sadiyah Afroz, Arvinpreet Kaur, Riddhi Sharma, Bhumika Chaudhary, and Diksha Marwaha**, who have not only contributed to my research journey but have also made it deeply meaningful on a personal level. Their unwavering support, respect, and care have touched me profoundly and have been a source of immense comfort during this journey. Thanks to Sadiyah, whose vibrant and bubbly personality, combined with her immense potential, has been truly inspiring. Her enthusiasm and creativity have always lifted my spirits, and the moments we shared have added so much joy and laughter to this experience. Thanks to Arvinpreet, for her humorous stories which have brightened my days and helped me believe in myself during difficult times. Your thoughtful gestures and the respect you have shown me have left a lasting impact. Riddhi, Bhumika, and Diksha, I deeply value the warmth, respect, and care you have always extended to me. Each of you has made a unique and cherished contribution to my life, and I am so grateful for the personal and professional bonds we share. Thank you for your presence, your support, and for always showing me such genuine respect and care.

I am deeply grateful to my friends from other labs who have been a constant source of support and joy throughout this journey:- **Ridam Pal, Subhadeep Duari, Vishakha Gautam, Sanjay Kumar Mohanty, Omkar Chandra, Aayushi Mittal, Sakshi Arora, Mansi Goel and Madhu Sharma**, each of you has played a special role in making this experience truly memorable. I sincerely thank Ridam for his invaluable support and guidance throughout my academic journey. As a senior from another lab, his technical expertise and advice have been instrumental in overcoming challenges and enhancing my work. Beyond professional support, his encouragement and understanding during difficult times have been a constant source of strength. I am deeply grateful for his kindness and unwavering help, both personally and professionally. Aayushi and Madhu, my batchmates from different labs, have shared similar struggles and challenges with me, and their companionship has made this journey much more manageable and fulfilling. Their constant support, thoughtful conversations, and shared moments of both frustration and joy have been a source of strength and encouragement. Subhadeep, though junior to me, shares a unique, child-like bond with me. While he respects me, he never misses an opportunity to tease me, and our camaraderie has brought countless moments of laughter and joy. Vishakha, Sanjay, Sakshi, and Mansi also juniors, have been incredibly supportive, offering valuable advice when I needed it

most, and we've shared numerous fun-filled moments together. Omkar, as a senior, has always been a source of thoughtful guidance and mentorship, which I deeply value. You have all made this journey unforgettable.

Finally, I would like to extend my heartfelt gratitude to all the colleagues and collaborators whose names I may have inadvertently missed. Each interaction throughout my research journey has left a meaningful impact, whether through engaging discussions, valuable feedback, or simply offering a listening ear. I deeply appreciate the unique contributions of everyone who has been a part of this journey and feel truly honored to have been part of such an inspiring and supportive community.

Gayatri Panda

ABSTRACT

KEYWORDS: CRISPR/Cas ; off-target activity ; specificity ; allostery; molecular dynamic simulations; accelerated molecular simulations; machine-learning.

CRISPR/Cas, a recently discovered genome-editing method depends on a single protein (Cas9) and non-coding RNA for gene-editing, which makes it simple, more rapid, versatile, efficient and manipulatable. Despite the high-efficiency and user-friendliness of CRISPR/Cas9, its applications are limited by various factors which includes large size of Cas9 (difficult in delivery), recognition to a specific PAM (limiting its effectiveness), differential specificity and sensitivity of Cas variants and introduction of random off-target mutations at sequences similar to those of target genes, major concern. An ortholog of Cas9 from Francisella novicida (FnCas9), was shown to have very low non-specific editing compared to SpCas9. The cleavage and recognition mechanism of FnCas9 as well as the reasons for its increased specificity have not yet been fully investigated, whereas SpCas9 (derived from Streptococcus pyogenes bacteria) has been the subject of much research. Questions pertaining to the molecular inter-play for the interactions and a comparison between the two orthologs can provide us insights into the elusive future of designing customized Cas9 molecules. This study provides a comprehensive examination of the structural and dynamic characteristics of SpCas9 and FnCas9, two prominent CRISPR/Cas9 orthologs, to elucidate the molecular mechanisms that govern their specificity, efficiency, and substrate versatility in genome editing. Utilizing molecular dynamics (MD) simulations, we compared the apo and gRNA-bound states of these proteins, revealing distinct conformational shifts and interactions that significantly influence gRNA binding and cleavage activity. FnCas9 demonstrated superior stability in gRNA binding attributed to dynamic domain rearrangements and specific residue interactions, particularly in the bridge-helix and REC3 domain. To further investigate specificity, accelerated MD simulations combined with machine learning techniques were employed to analyze RNA:DNA hybrid mismatches, revealing that FnCas9 maintains structural integrity in the presence of PAM-distal mismatches, unlike SpCas9. Enhanced interactions within the REC3 domain and allosteric communication pathways bypassing the REC2 domain were identified as critical factors in base-pair mismatch recognition and efficient gene-editing. Additionally, structural analyses of engineered

FnCas9 variants (en1, en15, en31) highlighted how specific mutations and domain modifications impact cleavage efficiency and specificity. The en31 variant exhibited distinct domain dynamics supporting better base-pair mismatch discrimination and adaptability for broader genome-editing targets due to its. Finally, the study compared FnCas9's interaction with RNA (tRNA) and DNA (tDNA) substrates, revealing unique interaction networks that suggest differing binding dynamics. While tDNA (target DNA) showed stronger binding affinity overall, tRNA (target RNA) binding induced greater conformational flexibility in key domains. These findings advance our understanding of the structural determinants of Cas9 function and provide actionable insights for engineering high-fidelity Cas9 variants with enhanced precision and substrate compatibility. Future directions include mechanistic studies on DNA-bound ternary complexes and the development of optimized Cas9 systems for RNA editing applications, contributing to the ongoing evolution of genome-editing technologies.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **i**

ABSTRACT **v**

LIST OF TABLES **xi**

LIST OF FIGURES **xx**

LIST OF PUBLICATIONS **1**

1 Introduction **2**

1.1 Genome Engineering using customisable molecular scissors 2

1.2 Development of CRISPR/Cas9 to a cutting-Edge Genome Editing Tool 3

1.2.1 CRISPR/Cas origin and discovery timeline 3

1.2.2 Evolution of Type-II CRISPR/Cas9 system for gene-editing 5

1.3 Structure and Domain Organization of CRISPR/Cas9 system 9

1.3.1 Structural arrangement of CRISPR/Cas9 system 9

1.3.2 Structural Comparison of Different States of SpCas9. 10

1.4 Applications and Challenges of CRISPR/Cas gene editing 13

1.5 Prediction and Mitigation of Off-target activity 18

1.5.1 **Strategies to mitigate off-target effects of CRISPR/Cas9** 19

1.6 FnCas9: The Future of Precise and Safe Genome Editing 25

1.7 Thesis Objectives and Organization 28

2 Background **33**

2.1 Molecular Dynamics and its Applications on the CRISPR/Cas9 System 33

2.1.1 Molecular Dynamics 33

2.1.2 Ensembles 34

2.1.3 Integration Algorithms 37

2.2 Molecular Dynamic Simulation of Proteins-Nucleic Acid Complexes 38

2.2.1 Force-Field 38

2.2.2 Boundary Condition 40
2.2.3 Enhanced Efficiency Methods 41
2.2.4 Enhanced Sampling Methods 41
2.2.5 MD Simulation Trajectory Analysis 43
2.3 Machine-Learning and it's application in understanding CRISPR/Cas9 Mechanism 44
2.3.1 Machine Learning Methods 45
2.4 Unraveling CRISPR-Cas9 Dynamics and Mechanisms Using Computational Approaches 48
2.4.1 Conformational Rearrangement of Cas9 endonuclease on Nucleic Acid Binding 50
2.4.2 Allosteric Mechanism of Cas9 protein in PAM recognition 52
2.4.3 Mechanism of target and cleavage by Cas9 54
2.4.4 Mechanistic insights into off-target effects 56
2.5 Conclusions 57

3 Comparative Structural and Dynamics Study of Free and gRNA-bound FnCas9 and SpCas9 Proteins 59

3.1 Introduction 59
3.2 Results and Discussions 65
3.2.1 Conservation Analysis reveals the diversity of Sp- and FnCas9 proteins. 65
3.2.2 gRNA-binding to Cas9 proteins stabilizes the complex. 65
3.2.3 Relative Domain Movement and Domain Structural Deformations are involved in the binding of gRNA to Cas9 proteins. 69
3.2.4 Larger domain deviations are observed in SpCas9 protein than in FnCas9. 73
3.2.5 Charge potentials reveal key insights into the greater specificity of FnCas9 protein. 75

4 Deciphering Cas9 Specificity: Role of Domain Dynamics and RNA:DNA Hybrid Interactions Revealed through Machine Learning and Accelerated Molecular Simulations 80

4.1 Introduction 80
4.2 Materials and Methods 82
4.2.1 Structural Modeling 82
4.2.2 Conventional MD simulations 83
4.2.3 Gaussian Accelerated MD simulations 84
4.2.4 MD Simulation Trajectory Analysis 86

4.2.5 Analysis of the RNA:DNA base-pair complementarity 86
4.2.6 Principal component analysis 89
4.2.7 Linear Interaction Energy 89
4.2.8 Markov State Modelling of Heteroduplex 89
4.2.9 Uniform Manifold Approximation and Projection Clustering of DNA Structural Properties and Free-Energy Landscape Analysis 90
4.2.10 Machine Learning modeling to classify heteroduplex states in SpCas9 and FnCas9 systems 91
4.2.11 MM/GBSA Calculations 95
4.2.12 Shortest Residue Interaction Pathway Identification 95
4.2.13 Statistical test 96
4.3 Results and Discussion 96
4.3.1 Diverse non-canonical base-pairing in the RNA:DNA hybrid enables mismatch tolerance 96
4.3.2 State-Dependent Dynamics Unveils Restrained Conformational Changes in FnCas9 Heteroduplexes 99
4.3.3 Enhanced mismatch-induced Cas9 domain deviations imply greater sensitivity and specificity 102
4.3.4 Mismatches affect hybrid recognition influencing Cas9's specificity . . 104
4.3.5 Shorter allosteric pathway ensures efficient information transfer enhancing specificity 108
4.4 Conclusion 111

5 Structural Insights into Mismatch Recognition by Engineered FnCas9 Variants 115
5.1 Introduction 115
5.2 Materials and Methods 116
5.2.1 Structural Modeling and Conventional MD simulations 116
5.2.2 Gaussian Accelerated MD simulations 118
5.2.3 MD Simulation Trajectory Analysis 118
5.2.4 MM/GBSA Calculations 119
5.2.5 Clustering Analysis 119
5.3 Results and Discussions 120
5.3.1 Conformational landscape and domain flexibility in engineered FnCas9 variants 120
5.3.2 DNA-Protein Interaction Analysis of en1FnCas9 and en31FnCas9 Variants 122

5.3.3 Cas9-hybrid binding dynamics and residue contributions 123
5.3.4 Enhanced off-target discrimination in engineered variants 125
5.4 Conclusions 127

6 Expanding FnCas9 Functionality: Comparative Analysis of RNA and DNA Targeting 130
6.1 Introduction 130
6.2 Materials and Methods 131
6.2.1 Structural Modelling 131
6.2.2 Classical MD simulations 132
6.2.3 MD Simulation Trajectory Analysis 133
6.2.4 MM/GBSA Analysis 133
6.2.5 Residue-specific RMSF Difference Analysis 133
6.2.6 Distance Analysis 134
6.3 Results and Discussions 134
6.3.1 Structural Stability and Flexibility of FnCas9 Complexes with tDNA and tRNA 134
6.3.2 Domain-Specific Motions and conformational changes in tRNA vs. tDNA Systems 134
6.3.3 Binding Affinity Differences and the Role of WED and PI Domains . . 136
6.3.4 Residue-Specific Binding Interactions: A Comparative Contact Map Analysis 137
6.3.5 Analysis of Residue-Specific Flexibility and Substrate-Specific Binding 138
6.3.6 Dynamic Distance Fluctuations of Key Binding Residues 140
6.4 Conclusions 140

7 Conclusions and Future Work 144
7.0.1 **Summary of Major Conclusions** 144
7.0.2 Future Work 148

LIST OF TABLES

3.1 Computational alanine scanning result for hotspot residues involved in binding of gRNA with SpCas9 and FnCas9 protein. All values are in kcal/mol. 77

4.1 Machine learning model performance of SpCas9 System using Protein, DNA and Protein+DNA features using RFE, Extra-Tree Classifier and BorutaSHAP feature selection methods 107

4.2 Machine learning model performance of FnCas9 System using Protein, DNA and Protein+DNA features using RFE, Extra-Tree Classifier, and BorutaSHAP feature selection methods 108

5.1 Binding affinity calculated between sgRNA-tDNA and Cas9-hybrid (domain-wise for HNH, RuvC, REC1, REC2, and REC3) for stable part of trajectory (200-250 ns) using MMGBSA method. 125

6.1 Binding energy comparison between FnCas9 and tRNA/tDNA substrates: The table shows the binding energies (kcal/mol) of FnCas9 with tRNA and tDNA substrates for a 20 bp region, as well as the interaction energies between the WED-PI domain and the substrate across three replicates. The data reveal a stronger binding affinity for the tDNA-bound complex compared to the tRNA-bound complex, with mean values reported ± standard deviation. 138

LIST OF FIGURES

1.1 A. Timeline highlighting major events of CRISPR development and gene therapy. 4

1.2 A. Natural vs. Engineered CRISPR systems. Natural CRISPR Pathway: 1. transcription of pre-crRNA and tracrRNA 2. binding of tracrRNA to pre-crRNA 3. cleavage of guide RNA from pre-crRNA 4. binding of inactive Cas9 nuclease to the guide RNA to produce the active Cas9 nuclease. B. Engineered CRISPR: 1. transcription of Guide RNA as a single sequence 2. Cleavage of target sequence by single guide RNA (sgRNA)-Cas9 complex. Double strand break is repaired either by the non-homologous end joining (NHEJ) or homology directed repair (HDR) mechanism. 6

1.3 A. CRISPR Locus organization in Type-II CRISPR Cas system comprising of an array of repetitive sequences (repeats, brown diamonds) interspaced by short stretches of non-repetitive sequences (spacers, colored boxes) and a set of CRISPR-associated (Cas) genes (colored arrows). B. Modular organization of CRISPR/Cas system: Small subunits (SS*) might be fused to the large subunit in several types I subtypes. The $*$ next to the CARF and HEPN domain labels indicates that other unknown sensor and effector domains can be involved in the signaling pathway. Dispensable genes are indicated by a dashed outline 8

1.4 A. Domain organization in typical SpCas9 structure. B. Crystal structure of SpCas9 (PDB ID: 4CMP) shown in cartoon representation. Individual Cas9 regions are colored differently. The HNH domain in yellow, REC lobe in green, RuvC domain in red, C-terminal domain in blue, Topo in light pink and bridge-helix is shown in violet color. C. Surface representation of SpCas9 depicting the two nucleic acid binding clefts on the molecular surface. 11

1.5 Structures of bound and unbound SpCas9s. A(i) Apo-Cas9 (PDB:4CMP) - NUC lobe (yellow) and REC lobe(limon) and Arg helix(blue) (ii) Cas9 binary complex(PDB:4ZT0) - NUC lobe (cyan), REC lobe (teal), and Arg Helix (magenta). (iii) Cas9 ternary complex - Cas9+sgRNA+tDNA (PDB:4OO8) ; NUC lobe (light pink), REC-lobe (deep purple), Arg helix (red), sgRNA (yellow), tDNA (green). B(i) Aligned structure of apo and pre-target bound states of Cas9(4CMP+4ZT0). (ii) Aligned structure of apo and ternary Cas9 (4CMP+4OO8). (iii) Aligned structure of binary and ternary Cas9s (4ZT0+4OO8). C (i)Apo form of Cas9 (4CMP). (ii). Pre-target bound state of Cas9 binary complex (4ZT0). (iii). Single-Stranded DNA bound state of Cas9 ternary complex(4OO8). D. Superposed images of NUC lobe (i) Apo and pre-target bound states of Cas9(4CMP+4ZT0), (ii) Apo and ternary Cas9(4CMP+4OO8). (iii) Binary and Ternary Cas9s(4ZT0+4OO8). 12

1.6 CRISPR/Cas9 system has been utilized to create disease-based models for various significant human diseases such as viral diseases, neurological diseases, cancer, ocular disease, blood diseases, cardiovascular diseases and disorders, and other complex genetic human diseases. 15

1.7 CRISPR/Cas9 system has been utilized to create disease-based models for various significant human diseases such as viral diseases, neurological diseases, cancer, ocular disease, blood diseases, cardiovascular diseases and disorders, and other complex genetic human diseases. 18

1.8 Off-target effects by CRISPR/Cas9 and strategies to mitigate it. 19

1.9 A. Domain architecture of Cas9 orthologs: Organisation of NUC lobe (HNH and RuvC domains) and REC lobe among different orthologs of Cas9 is shown. RuvC-I, II, and III (deep yellow, mustard yellow, and dark green), HNH domain (dark-grey), BH- Bridge Helix (grey), and REC lobe (green color). The PDB ID and species name of each ortholog (right side) and length of each domain are written in each box. B. Surface electrostatic potential map of REC lobes of Cas9 orthologs colored from −5kT/e (red) to +5kT/e (blue). 28

1.10 A. Structural and sequence comparison of SpCas9 and FnCas9. (A) Domain architecture of SpCas9 (Streptococcus pyogenes Cas9) and FnCas9 (Francisella novicida Cas9), showing the key functional domains and their residue boundaries. . 29

1.11 Thesis Organisation: The thesis comprises seven distinct chapters, including the introduction chapter, a chapter on the background and related literature, four chapters on each objective designed for deciphering the key structural determinants of CRISPR/Cas9, last chapter talks about the conclusion and future work. 32

2.1 In Periodic boundary consitions, the simulation box is surrounded by identical copies of itself in all directions (x, y, and z axes). Whenever a molecule in the original box moves out of one side, an equivalent molecule enters the box from the opposite side. 40

2.2 The workflow of metastable state classification showing input feature sets, labels and steps followed for prediction. A – Extraction of Input features; B- State identification using MSM. Using states as labels. C- Three features used for prediction of states in both Cas9s (a) RNA:DNA hybrid features, (b) Only Protein features , and (c) both. D. Steps followed for classification 48

2.3 A. Pre-catalytic to catalytic conformational transitions in the HNH domain. In cyan and magenta, the HNH domain and flanking linker regions (L1 and L2) are shown, respectively. The Protospacer Adjacent Motif is shown in dark red(Palermo (2019*a*)). 53

2.4 A. SpCas9 is bound to a gRNA and a tDNA (4UN3.pdb) architecture with a close-up view of the PAM recognition region. B. Residues are shown in sticks representation in which alanine mutations were introduced in East et al. C. At the essential edges of the allosteric pathway (Tyr771–Lys775 and Glu584–Arg905) (grey spheres) connection between the L1 (residues 765–780) and L2 (residues 906–918) loops (magenta) linking HNH (776-905) (green) and RuvC (red) domains is shown. D. "Shortest pathways" linking pairs of catalytic residues of the RuvC and HNH domains of the Cas9 with PAM. 54

2.5 A. Snapshots of Cas9 in action-(i) Checkpoint state: Loading of sgRNA into SpCas9. At this state, the HNH domain remains disordered while the enzyme investigates complementarity between the sgRNA and tDNA. (ii) Post catalysis conformation: The REC2 domain dissociates from the tDNA strand and the HNH domain undergoes a conformational rearrangement placing the catalytic residues in line with the phosphodiester backbone. (iii) Product State: SpCas9 back to checkpoint state with ordered REC2 domain and tDNA and ntDNA held by the REC2-REC3 domains and RuvC domain, respectively. These interactions are important for the stability of Cas9 bound to tDNA. B. QM/MM partitioning of the RuvC active site in CRISPR/Cas9. The QM part includes the Mg2+ ions surrounded by residues coordinating the metal ions. C. System's multiple competitive interactions are depicted in this diagram: Positively charged residues (blue color) at the boundary between HNH (yellow) and RuvC (pink) domain have electrostatic interactions with the negatively charged ntDNA (dark red) strand that keeps the ntDNA unwounded. 58

3.1 Figure: A. [Left] RMSD of backbone atoms over time for Apo-SpCas9 (violet) and RNA-bound SpCas9 (light-coral). [Right] RMSD probability distribution for Apo-SpCas9 and RNA-bound SpCas9. B. [Left] Radius of Gyration evolution for Apo and RNA-bound SpCas9. [Right] Probability distribution of Radius of Gyration for Apo-SpCas9 and RNA-bound SpCas9. C. [Left] Solvent accessible surface area (SASA) over time for Apo-SpCas9. [Right] SASA probability distribution for Apo and RNA-bound SpCas9. D. [Left] Time evolution of intra-protein hydrogen bonds for Apo-SpCas9 and RNA-bound SpCas9. [Right] Probability distribution of intra-protein hydrogen bonds for Apo and RNA-bound SpCas9. E. [Left] RMSD over time for Apo-SpCas9 (violet) and RNA-bound SpCas9 (light-coral). [Right] RMSD probability distribution for both. F. [Left] Radius of Gyration over time for Apo and RNA-bound SpCas9. [Right] Probability distribution of Radius of Gyration for both. G. [Left] SASA over time for Apo-SpCas9. [Right] SASA probability distribution for Apo and RNA-bound SpCas9. H. [Left] Intra-protein hydrogen bonds over time for Apo and RNA-bound SpCas9. [Right] Probability distribution of intra-protein hydrogen bonds for both. 66

3.2 A. Grouped Bar plot showing the percentage of conserved residues (residues with color codes $\geq$8) in SpCas9 (PDB ID: 4CMP) (on the left) and FnCas9 (PDB ID: 5B2Q) (on the right) domains and the contribution made by charged, uncharged and non-polar residues in conservation. B. Residue and domain-wise conservation of residues in SpCas9 (on the left) and FnCas9 (on the right). The color schemes followed for domains and conservation color codes are added to the color bar on the top. C. Average geometrical properties calculated from stable part of MD trajectories for SpCas9 and FnCas9 systems. 67

3.3 Comparison between the structural flexibility of Apo-SpCas9 and SpCas9 with gRNA. (A) Per-residue RMSF profiles were calculated from MD trajectories of Apo-SpCas9 (violet) and SpCas9+RNA (coral). The residues corresponding to every domain in the SpCas9 structure are shown. (B) The structural representations of Apo-SpCas9 and SpCas9+RNA mapped with per-residue RMSF values generated using UCSF Chimera. The color ranges from blue to red and denotes that RMSF varies from the lowest to the highest values. 68

3.4 Comparison between the structural flexibility of Apo-FnCas9 and FnCas9 with gRNA. (A) Per-residue RMSF profiles were calculated from MD trajectories of Apo-FnCas9 (Indian-red) and FnCas9+RNA (teal). The residues corresponding to every domain in the FnCas9 structure are shown in the top horizontal bar. (B) The structural representations of Apo-FnCas9 and FnCas9+RNA mapped with per-residue RMSF values generated using UCSF Chimera. The color ranges from blue to red and denotes that RMSF varies from the lowest to the highest values. 69

3.5 Domain Structural movements and transitions in Clustermap shows domain movement between domains on the X and Y-axis. Red boxes suggest domains moved closer and blue boxes are for domains that moved away. The color intensity depends on the magnitude of domain movement. (A) Clustermap for SpCas9 (B) Clustermap for FnCas9 Grouped bar plot with average RMSF values (200-1000ns) corresponding to RuvC, REC1, REC2, REC3, and HNH domains in the free and bound form of (C) SpCas9 system and (D) FnCas9 system. 71

3.6 Secondary structure and propensity analysis of Apo- and RNA-bound SpCas9 and FnCas9. A (i) and B (i) show residue-wise and domain-wise secondary structure assignment (present in >70% of frames in stable parts) for Apo-SpCas9 and SpCas9-RNA, respectively. A (ii) and B (ii) display secondary structure propensity per residue across 1000 frames for Apo-SpCas9 and SpCas9-RNA, with Helix (Brown), Sheet (Blue), and Coil (Yellow) probabilities. High fluctuation regions (RMSF) are boxed in red. C (i) and D (i) show the residue-wise and domain-wise assignment for Apo-FnCas9 and FnCas9-RNA. C (ii) and D (ii) display secondary structure propensity per residue across 1000 frames for Apo-FnCas9 and FnCas9-RNA with Helix (Violet), Sheet (Pink), and Coil (Cyan) probabilities. 72

3.7 Domain Structural movements and transitions in FnCas9. Domain-wise percentage of dynamic residues ($>$ 70% part of trajectory) in (A) Apo-SpCas9 and SpCas9 with gRNA and in (B) Apo-FnCas9 and FnCas9 with gRNA. 2D projection of backbone atoms in essential subspace along first two eigenvectors of (C) Apo-SpCas9(violet) and SpCas9 with RNA (lightcoral) (D) Apo-FnCas9 (indian-red) and FnCas9 with RNA (teal). Most dominant mode, containing 100 conformations along eigenvector 1 for (E) Apo-SpCas9 and (F) SpCas9 with gRNA (G) Apo-FnCas9 (H) FnCas9 with gRNA form and visualized in NMWiz of VMD. The arrows indicate the direction and relative amplitude of motions. 74

3.8 A. Free energy landscape generated using principal components (PC1 and PC2) for Apo-SpCas9 and SpCas9 with gRNA during 1000ns simulations.(B) Free energy landscape generated using principal components (PC1 and PC2) for Apo-FnCas9 and FnCas9 with gRNA during 1000ns simulations. Energy distribution is shown by the coloring pattern: Blue defines the conformational space with minimum energy (stable state) while red defines a conformational space with maximum energy (unstable state). Principal components are displayed as a contour map at the bottom of each FEL plot with similar color pattern like the energy landscape. . . . 76

3.9 A. Bar plots indicating the free energy components (vdW, electrostatic, polar surface, SASA, obtained from MM-PBSA calculations for gRNA binding with SpCas9 (left) and FnCas9 (right). B. Box plot showing the distribution of per-residue energy contribution in the binding of gRNA with SpCas9 and FnCas9. C. Electrostatic potential of HNH, RuvC, REC domains, and bridge-helix in SpCas9 and FnCas9. D. Hotspot residues involved in gRNA binding shown in sticks representation in SpCas9 and FnCas9. 78

4.1 (A) RMSD of SpCas9 backbone atoms for native and off-target forms with RMSD probability distributions; (B) RMSD of FnCas9 backbone atoms for native and off-target forms with probability distributions; (C-D) RMSD of RNA hybrid atoms for SpCas9 and FnCas9, respectively, with probability distributions; (E-F) Radius of gyration of SpCas9 and FnCas9 protein atoms with probability distributions; (G-H) Solvent Accessible Surface Area (SASA) of SpCas9 and FnCas9 with probability distributions for native and off-target forms. 87

4.2 Distribution (as boxplot) of the number of native contacts formed between ntDNA:tDNA in all the forms of (A). SpCas9, (B). FnCas9. (C). Pair-wise distribution (as violin plot) of the number of native contacts formed between sgRNA:tDNA in all the forms SpCas9 and FnCas9. (D) Pair-wise distribution (as boxplot) of the loss of RNA:DNA contacts formed in all the off-target forms SpCas9 and FnCas9. . . 88

4.3 Heatmap showing correlation among the features found in SpCas9 and FnCas9 94

4.4 Schematic diagram: Molecular Dynamic Simulation Setup. (1) On-target and off-target structural models used for performing the simulations. (2) Post-simulation trajectory analysis conducted for analysing the structural changes in Cas9 protein and RNA:DNA heteroduplex. Construction of the Markov state model (MSM). (1) Intra-base pair contacts at the distal end of the PAM were used for feature selection. (2) Identification of reduced key features using PCA dimension reduction method. (3) Clustering similar conformations in the trajectory using K-means. (4) MSM estimation and validation using the CK test. (5) The Transition probability calculation for every state is visualized as a graph, where the nodes represent each MSM state, and the thickness of the directed edges is the transition probability from two states. Metastable State Prediction. (1) RNA:DNA hybrid and Protein descriptors were extracted from trajectories to predict heteroduplex states (given by MSM). (2) Split the data into training (80%) and testing dataset (20%). (3) Feature Selection used for improving the prediction: RFE, BorutaSHAP, and Extra-Tree Classifier. (4) Model Building and Model Testing. (5) Evaluating the model's performance based on Accuracy, Precision, F1-score and Recall. Analysis using structural descriptors. (1) Calculation of the binding affinity of the REC and NUC lobe with the heteroduplex using the MM/GBSA method. (2) Density distribution of binding energies in SpCas9 and FnCas9 systems. (3) Per-residue energy decomposition analysis to find key residues involved in Hybrid and REC/NUC lobe binding in every state (X-axis: Residues; Y-axis: Energy kcal/mol). (4) Comparison of the allosteric pathway among the precleavage and checkpoint state of SPCas9 with the precleavage state of FnCas9. 97

4.5 A. Diagram of the sgRNA (cyan color) and tDNA (blue color) heteroduplex in CRISPR/Cas9, showing the location of base pair mismatches (mm) at PAM distal sites. The model systems considered for Gaussian accelerated MD (GaMD) simulations including the on-target DNA sequence (on-target) and DNA sequences containing base pair mismatches at PAM distal sites are shown. (i.e., mm20, mm16_19, and mm17–20). Geometrical base-pair descriptors used for explaining the DNA dynamics are shown. B. Principal Component Analysis (PCA) graphs illustrating the conformational distributions of the sgRNA:tDNA heteroduplex in on-target and off-target forms of SpCas9 (top) and FnCas9 (bottom). C. Stacked bar plot showing the distribution of non-canonical base pairs found in off-target systems of SpCas9 (left) and FnCas9 (right). D. The Cas9 ternary complex is shown as isosurface representations that represent the angle between PAM-proximal and PAM-distal duplexes (θ). Violin plot showing the distribution of angle in on-target and off-target systems of SpCas9 and FnCas9. E. Two-dimensional uniform manifold approximation and projection (UMAP) generated using only geometric base pair descriptors for the SpCas9 (top) and FnCas9 (bottom) systems. . . . 100

4.6 A. A 2D representation of the five metastable macrostates found in SpCas9, namely Sp-S0, Sp-S1, Sp-S2, Sp-S3, and Sp-S4; and four metastable macrostates in FnCas9, namely Fn-S0, Fn-S1, Fn-S2, and Fn-S3, generated by the RNApdbee tool Zok *et al.* (2018) using representative structures of each input state. B, C. MSMs illustrating the kinetic network for the SpCas9 and FnCas9 systems. Circles correspond to the macrostates, labeled by the macrostate codes reported in Figure A. The red circle corresponds to the most unstable state with the highest percentage of base-pair loss. The proportion of every macrostate is shown inside the circle. Structural representation of every state is provided inside the circle where the color represents the atomic fluctuations found in that state, ranging from 0.6 (gray) to 1.1 (blue). The gray arrows between macrostates indicate the presence of a probability flux between this pair of states at equilibrium, and the arrow thickness is proportional to the flux. Large gray arrows correspond to kinetically favorable transitions. Pie charts corresponding to every state represent the percentage composition of on-target and off-target forms of SpCas9 and FnCas9 in each state. Each on-target and off-target form of Cas9 is color-coded. 102

4.7 A. Average Protein Root Mean Square Fluctuations (RMSF) in angstroms (Å) calculated for the stable part of the trajectory (300-500 ns) representing the RuvC, HNH, REC1, REC2, REC3, and PI domains in both SpCas9 (top) and FnCas9 (bottom) on-target and off-target systems. B. Average Protein Radius of Gyration (Å) within the stable segments of the trajectory (300-500 ns) corresponding to the RuvC, HNH, REC1, REC2, REC3, and PI domains in SpCas9 (top) and FnCas9 (bottom). Data is presented for both on-target and off-target systems. C. Principal Component Analysis (PCA) graphs illustrating the conformational distributions of the HNH, RuvC, and REC2 domains in the on-target and off-target systems of SpCas9 (top) and FnCas9 (bottom). D. Two-dimensional Uniform Manifold Approximation and Projection (UMAP) generated using only protein features for SpCas9 (top) and FnCas9 (bottom) systems 105

4.8 Difference plot showing difference in (protein+hybrid) feature importance obtained using Extra-Tree classifier and BorutaSHAP method. Threshold based on 1.5 times the standard deviation from the mean of the difference in importance scores. Threshold for Extra-Tree method: +/- 0.032 and BorutaSHAP method: +/- 2.485. Red bars represent features that are more important for FnCas9, while blue bars represent features more important for SpCas9. Key features based on threshold are highlighted using yellow dot. 109

4.9 A. Two-dimensional projection of Uniform Manifold Approximation and Projection (UMAP) components, for SpCas9 (left) and FnCas9 (right) systems using RNA:DNA hybrid and Protein structural parameters. B. UMAP clusters obtained for StCas9 (*Streptococcus Thermophilus*) using only DNA, only protein and both DNA & protein features. C. Upset plot depicting the intersecting features identified through Recursive Feature Elimination (RFE), Extra-Tree Classifier, and BorutaSHAP feature selection methods in SpCas9 and FnCas9. D. Bubble plot showing feature importance scores of SpCas9 (red) and FnCas9 (blue) using the BorutaSHAP algorithm for intersecting selected features. E. Bubble plot showing feature importance scores of SpCas9 (red) and FnCas9 (blue) using extra-tree classifier method for all features. 110

4.10 A. Distribution of binding affinity between nuclease (NUC)-sgRNA:tDNA heteroduplex and recognition (REC)-sgRNA:tDNA heteroduplex found in SpCas9 and FnCas9 system. B. Key residues involved in the binding of heteroduplex with REc and NUC lobes in SpCas9 and FnCas9 proteins are depicted in stick representation. Color representation of sgRNA-cyan; tDNA-blue; HNH residues-yellow, RuvC residues-pink; REC3 residues-green, REC2-residues-beige color. C. Residue interaction network found between REC3-REC2-HNH domains shown in SpCas9 - checkpoint and pre-cleavage states and FnCas9 ternary complex. The recognition and catalytic sites of REC3 and HNH domains are shown using red spheres. These sites were used as source and sink residues to find the shortest path. REC3, REC2, and HNH domains are shown in dark green, beige and yellow colors in tube representation. The residues involved in the shortest path are shown in the close-up view as blue spheres. 114

5.1 Crystal structure of the F. novicida and it's engineered variants bound to a guide RNA, the target DNA, and nontarget DNA strand. A. FnCas9 (native), B. en1FnCas9 (E1369R), C. en31FnCas9 (G1243T, E1369R, E1449H) 117

5.2 A. Diagram of the sgRNA (cyan color) and tDNA (blue color) heteroduplex in CRISPR/Cas9, showing the location of base pair mismatches (mm) at PAM distal sites. Model systems considered for Gaussian accelerated MD (GaMD) simulations including the on-target DNA sequence (on-target) and DNA sequences containing base pair mismatches at PAM distal sites are shown. (i.e., mm20, mm16 19, mm17–20, and mm2). 120

5.3 A. Trajectory map difference between FnCas9 native and en1FnCas9. REC1 region showing higher shifts in en1FnCas9 (blue) B. Trajectory map difference between FnCas9 native and en31FnCas9. HNH and REC1 region showing higher shifts in en31FnCas9 (blue) C. Residue-wise RMSF line plot for FnCas9 (native-violet), en1FnCas9 (teal), and en31FnCas9 (pink). 121

5.4 A. Residue-wise RMSF for different Cas9 variants (i) FnCas9, (ii) en1FnCas9, (iii) en31FnCas9, (iv) en1ΔREC2, and (v) en31ΔREC2. B. Principal Component Analysis (PCA) graphs illustrating the domain-wise conformational distributions of the (i) Hybrid, (ii) REC1, (iii) HNH, (iv) REC2, (v) RuvC and (vi) REC3 in the native (violet), en1FnCas9 (teal), en31FnCas9 (pink), en1ΔREC2 (seagreen), and en31ΔREC2 (purple). 123

5.5 Contact plot visualizing the interaction between DNA nucleotides (depicted in pink) and protein residues from different domains of the en1FnCas9 and en31FnCas9 variants, both in their full-length and REC2-truncated forms. The protein residues are color-coded based on their domains, such as REC1 (teal), REC3 (light teal), RuvC (yellow), HNH (purple), BH (blue), WED (green), PI (orange), and the I region (lime). 124

5.6 Energetic contribution of residues involved in Cas9 and hybrid in A (i)FnCas9(native), B (i) en1FnCas9, and C(i) en31FnCas9. Contact network between Protein-sgRNA, and Protein-tDNA. Contacting residues (within 4Åin more than 80% frames) from different domains are coloured differently A(ii) Contact network in FnCas9 (native), B(ii) Contact network in en1FnCas9, C(iii) Contact network in en1FnCas9. 126

5.7 Stacked bar plot showing the distribution of non-canonical base pairs found in off-target systems of (i) FnCas9 (ii) en1FnCas9, and (iii) en31FnCas9. B. Distribution plot showing loss of base-pair count in off-target forms of (i) FnCas9 (ii) en1FnCas9, and (iii) en31FnCas9. 128

6.1 (A) RMSD plot (Angstorms) of protein backbone atoms over time for FnCas9-sgRNA-tRNA bound form (i) Line plot showing RMSD for three replicates. (ii) Probability distributions of RMSDs in all three replicates. (B) RMSF comparison in all three replicates of FnCas9-sgRNA-tRNA bound forms (i) Residue-wise RMSF line plot (Angstorms) (ii) Domain-wise-RMSF bar-plots in all three replicates. (C) RMSD plot (Angstorms) of protein backbone atoms over time for FnCas9-sgRNA-tDNA bound form (i) Line plot showing RMSD for three replicates. (ii) Probability distributions of RMSDs in all three replicates. (D) RMSF comparison in all three replicates of FnCas9-sgRNA-tDNA bound forms (i) Residue-wise RMSF line plot (Angstorms) (ii) Domain-wise-RMSF bar-plots in all three replicates. 135

6.2 Principal Component Analysis (PCA) graphs illustrating the domain-wise conformational distributions of the (A) RuvC, (B) HNH, (C) WED, (D) REC1, (E) REC2, (F) REC3, and (G) PI domains, (H) Domain-wise radius of gyration in all the replicates of FnCas9 bound to tRNA (shades of pink) and tDNA (shades of blue), ... 137

6.3 Domain-wise contact networks, highlighting key residues within different domains such as REC1, REC2, RuvC, HNH, WED, PI, and L1 and L2 loops. Residue-residue contact maps for different replicates of (A) FnCas9 bound to tDNA, (B) FnCas9 bound to tRNA, and (C) Venn-diagram showing number of common/unique interacting residues in all three replicates of (i) FnCas9 bound to tDNA (ii) FnCas9 bound to tRNA, and (iii) Both FnCas9 bound to tDNA and tRNA. 139

6.4 RMSF Analysis for Common and Unique Binding Residues in FnCas9-tDNA and Fncas9-tRNA systems: A. The central Venn diagram identifies residues common to both complexes (gray overlap), residues unique to tDNA binding (yellow), and residues unique to tRNA binding (blue). RMSF across three replicates (rep1, rep2, rep3) for residues involved in B. Specific binding to tDNA (left), C. Common in both tDNA and tRNA systems, and D. Specific binding to tRNA (right). RMSF difference plots made by subtracting RMSF of residues in tRNA from tDNA system. Positive value (blue bar) means higher RMSF in tDNA system, and vice-versa (red bar). E. RMSF difference plot for residues unique to tDNA. F. RMSF difference plot for residues unique to tRNA, G. RMSF difference plot for residues common in both tDNA and tRNA. 141

6.5 Distance plot for residues (1082, 857, and 858) unique to tRNA bound system which went farther away in tDNA system for (A) Replicate 1, (B) Replicate 2, and (C) Replicate 3 of both tRNA and tDNA bound systems. 142

6.6 Distance plot for residues (845, 848, and 1042) unique to tDNA bound system which went farther away in tRNA system for (A) Replicate 1, (B) Replicate 2, and (C) Replicate 3 of both tRNA and tDNA bound systems.) 143

LIST OF PUBLICATIONS

Thesis-Related Publications

- Panda G, Ray A. Deciphering Cas9 Specificity: Role of Domain Dynamics and RNA:DNA Hybrid Interactions Revealed through Machine Learning and Accelerated Molecular Simulations. *Int. J. Biol. Macromol.*, 2024. https://doi.org/10.1016/j.ijbiomac.2024.137835

- Panda G, Ray A. Comparative structural and dynamics study of free and gRNA-bound FnCas9 and SpCas9 proteins. *Comput Struct Biotechnol J.*, 2022 Aug 4;20:4172-4184. https://doi.org/10.1016/j.csbj.2022.07.041

- Panda G, Ray A. Decrypting the mechanistic basis of CRISPR/Cas9 protein. *Prog Biophys Mol Biol.*, 2022 Aug;172:60-76. https://doi.org/10.1016/j.pbiomolbio.2022.05.001

Other Publications

- Ray S, Gupta S, Panda G, Chatterjee P, Das A, Patawri P, Hosseinzadeh P, Ray A, Banerjee S. Identification of pseudobaptigenin as a novel polyphenol-based multi-target antagonist of different hormone receptors for breast cancer therapeutics. *J Biomol Struct Dyn.*, 2023 Jul;42(11):5502-5514. https://doi.org/10.1080/07391102.2023.2226750

- Panda G, Mishra N, Sharma D, Kutum R, Bhoyar RC, Jain A, Imran M, Senthilvel V, Divakar MK, Mishra A, Garg P, Banerjee P, Sivasubbu S, Scaria V, Ray A. Comprehensive Assessment of Indian Variations in the Druggable Kinome Landscape Highlights Distinct Insights at the Sequence, Structure and Pharmacogenomic Stratum. *Front Pharmacol.*, 2022 Jul 5;13:858345. https://doi.org/10.3389/fphar.2022.858345

- Haider M, Anand V, Enayathullah MG, Parekh Y, Ram S, Kumari S, Anmol, Panda G, Shukla M, Dholakia D, Ray A, Bhattacharyya S, Sharma U, Bokara KK, Prasher B, Mukerji M. Anti-SARS-CoV-2 potential of *Cissampelos pareira L.* identified by connectivity map-based analysis and in vitro studies. *BMC Complement Med Ther.*, 2022 Apr 22;22(1):114. https://doi.org/10.1186/s12906-022-03584-3

- Panda, G., Mishra, N., & Ray, A. Genetic variations and drug repurposing provide key insights into the disruption of the SARS COV2. (preprint). 2020. https://doi.org/10.31219/osf.io/b7y2c

- Mishra, D., Panda, G., Kumar, P., & Singh, S. Novel drug delivery system for herbal formulation in cancer treatment. *World J Pharm Res.*, 2017;6:341-353. https://doi.org/10.20959/wjpr201715-10093

CHAPTER 1

Introduction

1.1 Genome Engineering using customisable molecular scissors

Genetic disorders are caused by one or more mutations in the genome, by a combination of genetic mutations and environmental factors, or by chromosomal damage, which could be treated by gene therapy or gene editing. Gene therapy does this by inserting a correct copy of the gene into the genome of cells in the target organ or tissue. In contrast, gene editing modifies the genome at a specific site to repair or alter the genetic sequence . Genome editing technology involves precise and efficient genetic alterations (add/remove) of particular locations in the genome to study a gene function, repair or replace a defective gene (gene therapy), thus correcting some genetic defects or applying it for economic benefits (crop improvement) Konermann *et al.* (2015); van der Oost *et al.* (2014). The applications of gene editing include the modification of specific genes and observation of their effects on organisms to study how they function. Gene editing can also be used to cure specific diseases through precise genetic modifications, improve crop yields and disease resistance, and gain a better understanding of complex cellular processes. Gene editing can be achieved by programming nucleases, such as zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) system (CRISPR-Cas9). These nucleases, introduce double-strand breaks in the DNA, which is prepared through homology-directed repair (HDR) or non-homologous end-joining (NHEJ).

Zinc-finger nucleases (ZFNs) enzymes are created by combining a DNA cleavage domain with a zinc-finger DNA-binding domain. The zinc finger domain, which is generally composed of about 30 amino acid residues, binds to three base pairs on each repeat of a DNA sequence to recognize it Kim and Kim (2014). The DNA cleavage domain is derived from the type II restriction endonuclease FokI, which must dimerize to cleave DNA, meaning that two ZFNs are required to cause double-strand breaks. It can target longer DNA stretches by assembling multiple zinc fingers. Although widely used for precise genomic alterations, assembling ZFNs to target a polynucleotide stretch is very challenging and time-consuming Urnov *et al.* (2010). To address these challenges researchers designed transcription activator-like effector nucleases (TALENs). It is secreted by Xanthomonas bacteria which can alter host plant transcription. TALENs were engineered to target specific genes by attaching TALEs to the FokI cleavage domain and then customizing the TALE repeats with specific repeat-variable diresidues (RVDs). TALENs and ZFNs have shortcomings despite their versatility. For cleavage, both require the FokI endonuclease to dimerize,

which increases complexity and may reduce targeting efficiency Costa *et al.* (2017); Singh *et al.* (2016). Furthermore, the expense of TALENs can be high because of the requirement for custom repeats. The clustered regularly interspaced short palindromic repeats (CRISPR) system is a newer and more customizable targeted nuclease that can address the limitations of ZFNs and TALENs. CRISPR has shown significant promise in gene editing and may provide a more effective and cost-efficient option.

1.2 Development of CRISPR/Cas9 to a cutting-Edge Genome Editing Tool

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and CRISPR-associated (Cas) gene-editing method was adopted from the natural-defense mechanism present in bacteria (40%) and archaea (90%) against invading virus and plasmids. Bacteria adapt immunity against past infections and record in the CRISPR locus as spacers to attain resistance against future infections. A typical CRISPR locus consists of a CRISPR array with 20-50 bp DNA repeats (brown diamonds in Figure 1.3A) interspersed by unique 'spacer' sequences (30-40 bp, colored boxes in Figure 1.3A) with CRISPR-associated genes, or Cas genes and sequence-specific RNA components. The spacers serve as a genomic record to recognize and generate an immune response against similar viral invasions in the future. Since the discovery of the CRISPR locus, numerous research groups have further demonstrated that the CRISPR-Cas system is an RNA-guided, adaptive immune system and have described the various mechanisms of actions along with methods for adopting Cas proteins in different biotechnological applications. In the following sections, the origin, discovery, and classification of CRISPR/Cas system, are summarized.

1.2.1 CRISPR/Cas origin and discovery timeline

CRISPR-Cas systems were discovered recently and quickly gained popularity, undergoing thorough characterization. The first observations of unusual repetitive structures surfaced 30 years ago, and there has been a surge in interest since 2012-2013, when CRISPR/Cas started being used for genome editing task Makarova and Koonin (2015); Jiang and Doudna (2015). The origin of CRISPR as an adaptive immune system in bacteria, the elucidation of CRISPR-Cas9 biological and molecular mechanisms, and the subsequent engineering and optimization of it as a programmable genome editing tool are remarkable achievements in science highlighted in Figure 1.1.

In 1989, Anderson's group conducted the first successful gene transfer in humans, correcting a genetic defect in patients with adenosine deaminase (ADA) deficiency. This pioneering work laid the groundwork for future gene therapy applications Sakiyama (1996). By 1990, the first gene delivery for therapeutic intent was performed in children with severe combined immunodeficiency (ADA-SCID), marking a critical step in translating gene therapy from concept to clinical

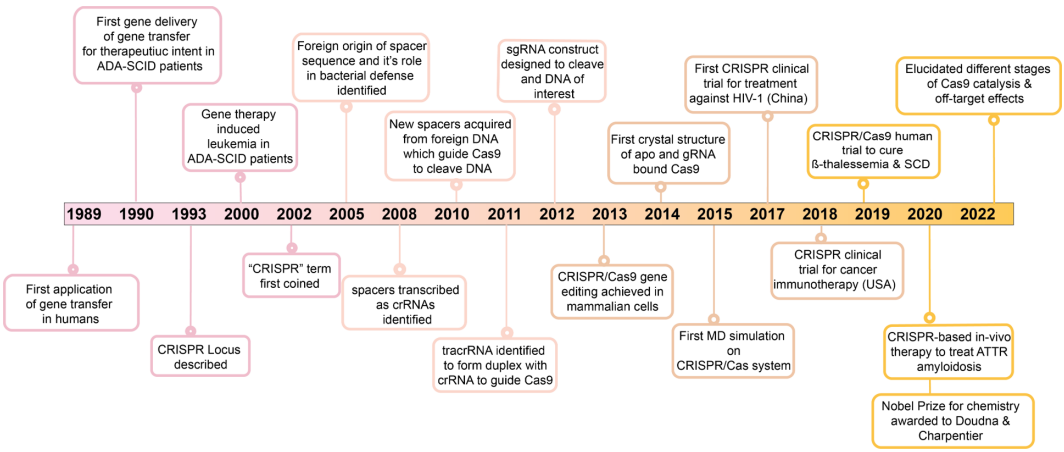


Figure 1.1: A. Timeline highlighting major events of CRISPR development and gene therapy.

application Blaese *et al.* (1995). However, the early optimism was tempered by setbacks; in 1993, leukemia was reported in some ADA-SCID patients, highlighting the risks associated with integrating viral vectors in gene therapy Howe *et al.* (2008). The understanding of CRISPR technology began to take shape in 2000 when Francisco Mojica and colleagues described the CRISPR locus in prokaryotes, revealing its role in bacterial immunity Mojica *et al.* (1993). The term "CRISPR" was first coined by Jansen et al. in 2002, formalizing the discovery of these sequences Jansen *et al.* (2002). Researchers soon started to uncover the specifics of how CRISPR combat invading phages. The initial crucial discovery was made by John van der Oost and his team, who demonstrated that E.coli that spacer sequences originate from phages are transcribed into small RNAs called CRISPR RNAs (crRNAs), which direct Cas proteins to the target DNA Brouns *et al.* (2008*a*). Development of CRISPR for genome editing accelerated in 2011 when Emmanuelle Charpentier and colleagues identified tracrRNA, which forms a duplex with crRNA to guide the Cas9 protein to its target DNA. In 2012, Jennifer Doudna and Charpentier fused tracrRNA and crRNA into a single guide RNA (sgRNA), enabling precise gene editing Deltcheva *et al.* (2011). The first crystal structure of Cas9 was elucidated in 2013, enhancing the understanding of its mechanism of action Jinek *et al.* (2014*b*). In 2014, CRISPR/Cas9 gene editing was successfully achieved in mammalian cells, marking a significant advancement in genetic engineering Cong *et al.* (2013*b*); Doudna and Charpentier (2014). The following years saw the application of CRISPR technology in clinical settings: in 2017, the first CRISPR clinical trial for HIV-1 was initiated in China, and in 2018, trials began for treating β-thalassemia and sickle cell disease (SCD) Lu *et al.* (2020). The potential of CRISPR technology was further demonstrated in 2019 with its application in cancer immunotherapy. First in-vivo CRISPR therapy aimed at treating transthyretin amyloidosis (ATTR) was reported in 2020, the same year that Doudna and Charpentier were awarded the Nobel Prize in Chemistry for their groundbreaking work on CRISPR-Cas9 Frangoul *et al.* (2021). Recent studies in 2022 have elucidated the different stages of Cas9 catalysis and characterized off-target effects, refining the CRISPR technology further.

1.2.2 Evolution of Type-II CRISPR/Cas9 system for gene-editing

The CRISPR-Cas9 system of Streptococcus pyogenes serves as a powerful adaptive immune mechanism against bacteriophage infections. This system is characterized by three key stages: adaptation, expression, and interference, shown in Figure 1.2. During the adaptation phase, fragments of DNA from invading phages, known as spacers, are integrated into the CRISPR array, which is a segment of the bacterial genome containing repetitive sequences interspersed with these unique spacers. This integration allows the bacterium to "remember" past infections. In the expression phase, the CRISPR array is transcribed into a long precursor RNA (pre-crRNA), which is subsequently processed into smaller CRISPR RNAs (crRNAs). These crRNAs are crucial as they guide the Cas9 protein to specific DNA sequences during the interference phase. The crRNA forms a complex with a trans-activating CRISPR RNA (tracrRNA), and together they associate with the Cas9 nuclease. The complex then binds to the target DNA, where the crRNA hybridizes with its complementary sequence, leading to cleavage of the DNA by Cas9. This double-stranded break (DSB) is generally 2-6 base pairs upstream of a tri-nucleotide sequence in the target DNA, the adjacent protospacer motif (PAM) Bhaya *et al.* (2011); Barrangou *et al.* (2007). In synthetic applications, the natural crRNA and tracrRNA can be replaced with a single guide RNA (sgRNA), simplifying the system and enhancing its utility in genome editing. The cleavage of the target DNA creates a double-strand break, which can be repaired by either non-homologous end joining (NHEJ) or homology-directed repair (HDR), allowing for precise genetic modifications. The dependency of the CRISPR/Cas system on a single non-coding RNA molecule to target DNA for site-specific cleavage makes it more straightforward, faster, versatile, adaptable, efficient, and manipulable compared to the methods above Costa *et al.* (2017); Li *et al.* (2020); Manghwar *et al.* (2019). Typically, NHEJ will successfully repair a double-strand break (DSB) produced by Cas9. Nevertheless, NHEJ is susceptible to mistakes and often results in insertions and deletions (indels) in the specific region that requires repair. A frameshift mutation occurs when indels that develop in the coding region of a gene result in the loss of its functioning. This phenomenon is referred to as a gene knockout (KO). Gene knockouts are used in a range of research areas, including functional genomics, pathway analysis, drug discovery and screening, and disease modeling. Using multiple guide RNAs that target various regions of the gene ensures high-efficiency gene knockout, a method that is gaining popularity. In the presence of a double-strand break (DSB) caused by Cas9, cells can repair themselves by homologous double-replication (HDR). This mechanism presents a potential avenue for researchers to introduce a new segment of DNA or an entire gene fragment. This technique is referred to as a gene knock-in reaction. The presence of a DNA template, referred to as a donor template, is necessary in order to accomplish a gene knock-in. This donor template comprises the sequence or gene of interest surrounded by homology sections that correspond to the area on either side of the cut. The donor is delivered to the cells being engineered, along with the other editing components of Cas9 and sgRNA.

Based on Cas-gene repertoire, sequence similarity, and CRISPR-locus architecture, the CRISPR/Cas system has been classified into two classes and further subdivided into six types and nineteen subtypes Makarova and Koonin (2015); Delhove *et al.* (2020). The effector complex involved in the

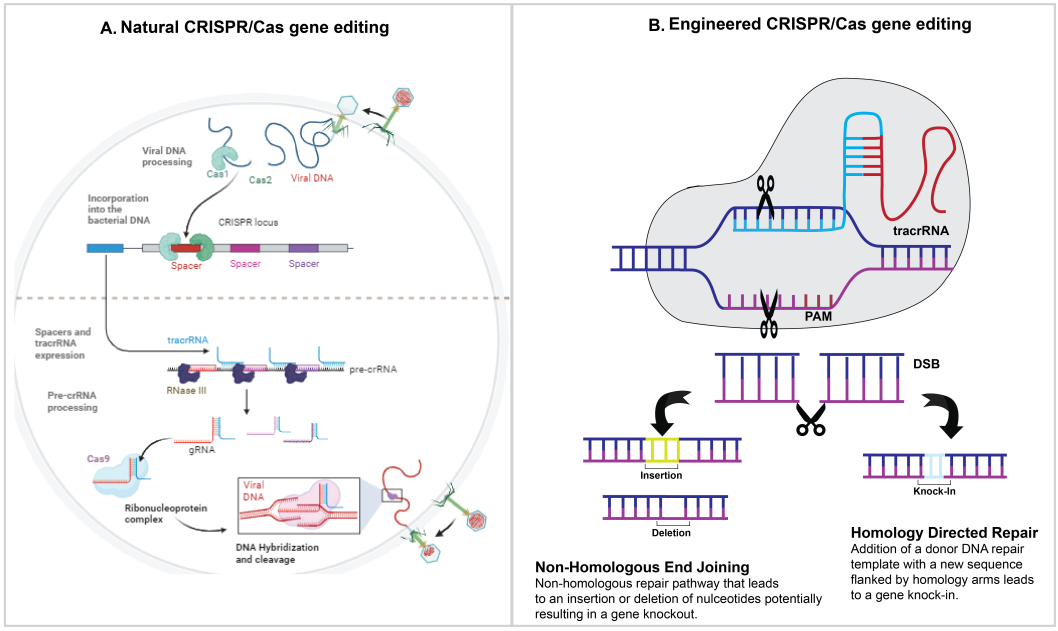


Figure 1.2: A. Natural vs. Engineered CRISPR systems. Natural CRISPR Pathway: 1. transcription of pre-crRNA and tracrRNA 2. binding of tracrRNA to pre-crRNA 3. cleavage of guide RNA from pre-crRNA 4. binding of inactive Cas9 nuclease to the guide RNA to produce the active Cas9 nuclease. B. Engineered CRISPR: 1. transcription of Guide RNA as a single sequence 2. Cleavage of target sequence by single guide RNA (sgRNA)-Cas9 complex. Double strand break is repaired either by the non-homologous end joining (NHEJ) or homology directed repair (HDR) mechanism.

maturation of the crRNAs and recognition of target DNA (tDNA) and cleavage is a multi-protein complex in Class 1 CRISPR/Cas systems. In contrast, a single Cas protein is involved for the same in the Class2 CRISPR/Cas system making it more favorable for applications in genome editing Costa *et al.* (2017). The phylogenetic analysis played a pivotal role in comprehending the development of Cas protein and determination of its functionally important regions, as well as it helped in the identification of novel Cas9 orthologs. Conserved key catalytic regions in Cas9 orthologs may indicate functionally essential residues. In 2006, Cas gene neighborhood studies were carried out on forty genomes, resulting in identifying twenty-five protein families Anderson *et al.* (2018). Due to the semi-independent evolution of various modules in CRISPR/Cas systems, multiple criteria were considered in the classification of this system, including the repertoire of signature Cas genes, structure and design of the Cas operons, and phylogeny of conserved Cas proteins. Three major types of CRISPR/Cas systems (I-A to I-F, II-A to II-C, and III-A to III-B) were reported in a study by Makarova and the team Makarova *et al.* (2011); Koonin and Makarova (2019). These three Cas forms were further divided into 11 subtypes, each identified by a signature gene. The expression and interference phases of the type-I and type-III Cas Systems are similar as they both require only crRNAs and Cas proteins. The type-II CRISPR/Cas system needs processing of tracrRNA and pre crRNA by RNAaseIII to produce dual- tracrRNA:crRNA, which directs the Cas9 endonuclease (Csn1) to the cleavage site Slaymaker *et al.* (2016*a*); Haft *et al.* (2005). Chylinski and group cataloged 235 homologous Cas sequences in 203 bacterial species Chylinski *et al.* (2013). In their analysis, truncated, incomplete, and highly similar sequences were removed,

and only 75 diverse representative sequences were analyzed, whereas 56 of them were predicted as novel Cas9 orthologs. The multiple sequence alignment of the Cas9 orthologs disclosed that there exists a higher diversity in their amino acid composition, tracrRNA gene position in CRISPR loci, and protein size (984 aa- C.jejuni, 1629aa F.novicida). However, they share the same domain architecture with a central HNH endonuclease domain and split RuvC/RNase H domain. The Cas9 orthologs were grouped into three subclusters with diverse type II CRISPR Cas loci architecture, II-A (Csn2-like protein), II- B (Csx12- like Cas9, and Cas4), and II-C (Cas9, Cas1, and Cas2 genes). Later, homology between Cas9 and nucleases encoded by mobile genetic elements (transposons) was identified, indicating that type-II CRISPR-Cas evolved via recombination of mobile nuclease genes with type-I loci Chylinski *et al.* (2014). In work published by Fonfara and colleagues Fonfara *et al.* (2014), a remarkable sequence diversity was observed in the evolution of dual-RNA (sgRNA) and Cas9 in bacteria. They selected eight representative sequences of phylogenetically defined type II CRISPR/Cas groups and studied the coevolution (modularity and variability) of Cas9 and dual-RNA. There were many interpretations made from their study, such as the evolution in Cas9-gRNA occurring both vertically and horizontally. The ability of Cas9 to specifically recognize different gRNAs is due to the presence of highly variable and less conserved inter-motif regions in members of various subgroups of Cas9. The findings made by Fonfara's research group expanded the number of Cas9 enzymes and enlightened the evolutionary behavior of type II CRISPR/Cas systems, dual-RNA structures, and associated complex PAMs. In 2015, Makaravo's group adopted a two-step classification approach for Cas genes. The first step involved the identification of all Cas genes in CRISPR/Cas locus, followed by determination of the signature genes and unique gene architecture for their sub-classification into types and sub-types. The classification of CRISPR/Cas system was done on the basis of effector molecules involved, multisubunit crRNA–effector complexes belonged to Class 1 systems, whereas single effector protein carried out all the functions, then it is classified as Class 2 Makarova and Koonin (2015); Makarova *et al.* (2015). In 2017, Koonin and group Koonin *et al.* (2017) identified new types and sub-types belonging to Class 2 CRISPR/Cas system. They discovered the difference in both CRISPR/Cas systems in their modular organization and observed that the effector module in Class 2 system involves large, multidomain proteins that also participate in pre-crRNA processing (Type V and VI). Figure 1.3A shows the modular organization of both CRISPR/Cas classes. Each class has three types, Class1 has types I, III and IV and types II, V and VI belong to Class 2. Every type has a unique architecture of the effector modules and signature proteins. The Figure 1.3A is inspired from a review published in 2019 by Makarova group depicting the modular organisation of both Class 1 and Class2 CRISPR/Cas along with the functional contribution of each subunit Koonin and Makarova (2019). A study was published Karvelis *et al.* (2019) in 2018, in which Karvelis and his coauthors have proposed a bioinformatic pipeline for the characterization of new Cas9 orthologs. The method for characterization was divided into three sections, identification of new Cas9 orthologs, followed by Cas9 expression and PAM identification. In the identification of CRISPR/Cas orthologs step, all the Type-II Cas9 endonucleases with CRISPR repeats were identified using PILERCR Edgar (2007), and MinCED Bland *et al.* (2007). The Cas9-like proteins were searched using hidden Markov models or position-specific scoring matrices developed for Cas9.

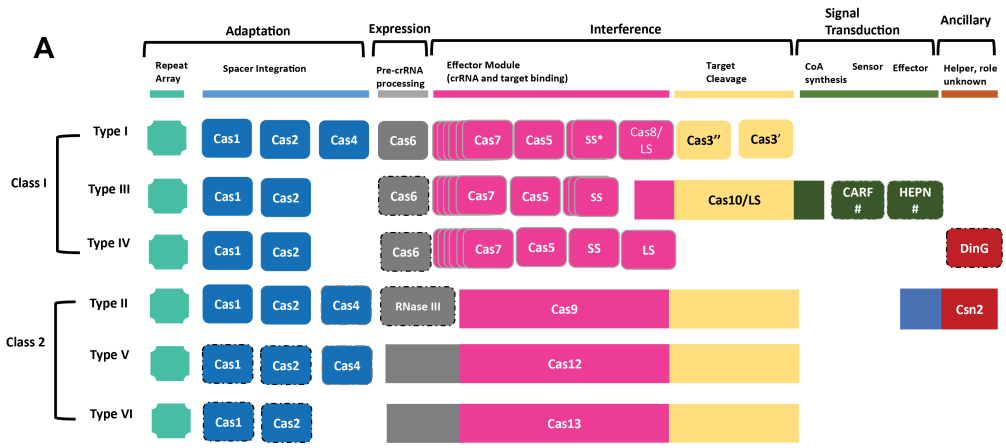


Figure 1.3: A. CRISPR Locus organization in Type-II CRISPR Cas system comprising of an array of repetitive sequences (repeats, brown diamonds) interspaced by short stretches of non-repetitive sequences (spacers, colored boxes) and a set of CRISPR-associated (Cas) genes (colored arrows). B. Modular organization of CRISPR/Cas system: Small subunits (SS*) might be fused to the large subunit in several types I subtypes. The $*$ next to the CARF and HEPN domain labels indicates that other unknown sensor and effector domains can be involved in the signaling pathway. Dispensable genes are indicated by a dashed outline

The presence of HNH or RuVC domain confirms the protein is Cas9 which can be accomplished by performing MSA or MSAPRobs, etc. Karvelis *et al.* (2019). Makarova and the group came up with an updated classification in 2019 that contained two classes, six types and thirty-three subtypes, compared with five types and sixteen subtypes in 2015. In the Class 2 system, 3 types and 17 Cas subtypes were newly discovered and along with it quite a few ancillary CRISPR-linked genes, involved in signal transduction were identified. Some novel Cas variants associated with transposons were also found that lacked nucleases required for interference. These discoveries add to our understanding of CRISPR/Cas functional variety and evolutionary history Makarova *et al.* (2020) (Figure 1.3A). The pipeline suggested by Karvelis and his coauthors in 2018 Karvelis *et al.* (2019) was followed by Gasiunas and his coworkers in their recent work published in 2020, in which they characterized 79 phylogenetically distinct Cas9s Gasiunas *et al.* (2020). Cas9 protein sequences were aligned using MUSCLE under default parameters, the alignment was exported in ClustalW format, later imported into Clustal Omega software, and a percent identity matrix was generated. The sequence alignment of these sequences had shown a variation of 93%. It was observed that within a particular phylogenetic clade of a Cas9 ortholog, the spatial positioning and orientation of the crRNA and tracrRNA were much conserved. The Cas9 orthologs and their metagenomic sequences were taken from a publicly available repository Gasiunas *et al.* (2020)). The findings from phylogenetic and clustering research mentioned in this section led to the characterization of unique orthologous Cas9 nucleases with distinct and potentially desirable features compared to those commonly given to SpCas9.

1.3 Structure and Domain Organization of CRISPR/Cas9 system

The most widely studied CRISPR/Cas system is the type IIA system where Cas9 is the only protein required for the interference step. It is a complex made up of single guide RNA (sgRNA; where crRNA and tracrRNA have been fused) and Cas9 protein, a 160 kDa DNA endonuclease enzyme that cleaves each strand of double-stranded DNA at a precise position.

1.3.1 Structural arrangement of CRISPR/Cas9 system

SpCas9 (Streptococcus pyogenes Cas9) protein, the most widely used molecular scissors for genetic engineering, was the first biochemically characterized Cas9 with a 2.6 Å resolution deposited by Jinek and colleagues in 2014 Jinek *et al.* (2014*a*). The SpCas9 enzyme is 1368aa long (4.1kb) and adopts a bilobed architecture containing a nuclease lobe(NUC) and alpha-helical lobe or Recognition lobe shown in Figure1.4. The Nuclease lobe(51-55 and 719-1368) comprise HNH (residues 780-906) and the RuvC-like domain residues(1-55, 719-765, and 919-1099),C-terminal domain(residues 1100-1368). The HNH and RuvC domains help in DNA cleavage by bringing double-strand breaks in the DNA. While the REC lobe consists of three alpha-helical domains (Hel-I, Hel- II, and Hel-III), none of which are structurally similar to other proteins. This lobe contains a' bridge helix (residues 59-76), which connects the NUC and REC lobes with positively charged arginine residues; alterations in residues in this lobe are responsible for the variation in Cas9 sizes seen among different orthologs Nishimasu *et al.* (2014*a*); Josephs *et al.* (2015*a*).

SpCas9's RuvC domain forms the structural core of the NUC lobe and is divided into three discontinuous segments (RuvC I-RuvC III), with the HNH domain lying between the second and third segments (Figure 1.4A). Each segment is made up of a six-stranded sheet surrounded by four helices. SpCas9's C-terminal domain (CTD) has a $\beta-\beta-\alpha-\beta)$ secondary structure arrangement similar to a typical Topoisomerase II structure Jinek *et al.* (2014*a*). The C-terminal domain is a mix of α/β regions ranging from 1201 to 1363 residues containing PAM recognition regions that are highly irregular in apo-Cas9 (PDB-IB 4CMP, Figure 1.4), keeping it inactive thereby preventing it from recognizing the target DNA before binding to a guide RNA. This structural finding is consistent with DNA curtains assays, which show that apo–Cas9 binds DNA non-specifically and can be quickly detached from non-specific sites in guide RNA or heparin Josephs *et al.* (2015*a*). The SpCas9 has a total surface area of 1034 $Å^2$ buried between the two structural lobes. SpCas9 consists of two nucleic acid binding clefts, a narrower one in the nuclease and a wider one in the recognition lobe sufficient to accommodate an RNA-RNA or DNA-RNA duplex. The cleft in the NUC lobe is 40 Å long, 15 to 20 Å wide, and 15 Å deep, and it is flanked by the active site of RuvC at the bottom, CTD on one side, and the HNH domain and REC lobe bulge on the other. The surface of the groove is lined by highly positively charged residues rich in Arginine (Arg69, Arg70, Arg71, Arg75, and Arg76). The high conservation of amino acids located in grooves of the

NUC and REC lobe indicates the importance of the function of these clefts. SpCas9 recognizes a 5'-NGG-3' PAM sequence located three base pairs (bp) upstream to the 3' side of the cleavage site on the non-target DNA strand, whereas other Cas9 orthologs have different PAM requirements.

The Research led by Jinek in 2014 shed light on Cas9's function, regulation, and evolution. The structures of type II-A and type II-C subtypes revealed the basic design of an RNA-guided DNA endonuclease with well-defined nucleic acid binding clefts. Biochemical experiments reported in the paper revealed that in the absence of sgRNA, the binding site responsible for PAM recognition is disordered. Single-particle electron microscopy (EM) structures had shown that Cas9 binding with guide RNA leads to the reorientation of two structural lobes of Cas9 Jinek *et al.* (2014*a*). Also, in the same work, it was reported that the conformational rearrangement of Cas9 was comparatively small on binding to tDNA and PAM sequence, indicating that significant structural changes occur prior to tDNA binding suggesting guide RNA loading as a critical step in Cas9 activation. In the same year, Nishimasu and the group deposited Nishimasu *et al.* (2014*a*) the 2.5Å resolution (PDB ID: 4OO8) crystal structure of a complex of SpCas9 bound to sgRNA and its target DNA. SpCas9's two lobes accommodate the sgRNA-DNA hetero-duplex in a positively charged groove at their interface, according to the structure. The REC lobe is required for sgRNA and DNA binding, while the nuclease lobe domains (HNH and RuvC) are necessary for cleavage of the target and non-target strands of DNA, respectively. The interaction with the protospacer adjacent motif is regulated by the NUC lobe's C-terminal domain (PAM). This study clarified the molecular mechanism of Cas9's RNA-guided DNA targeting. The structural discoveries discussed here establish the groundwork for understanding how Cas9 detects and cleaves target DNA, as well as for future studies into target selectivity and cleavage efficacy.

1.3.2 Structural Comparison of Different States of SpCas9.

Palermo and group, 2017, have talked about the structural transitions of Cas9 protein on binding with RNA and DNA. The conformational transitions of Cas9 from the "open" apo state to a "closed" conformation, suggestThe conformational transitions of Cas9 from the "open" apo state to a "closed" conformation suggest a mechanism for RNA recruitment in which domain reorganizations cause the formation of a positively charged cavity for nucleic acid binding.e active state that mediates in the DNA cleavage process. A comparative analysis of Cas9 states can aid in understanding the association of the structural changes with the cleavage process Palermo *et al.* (2017). The structural comparison of different states of SpCas9 crystal structures representing the free state (apo-Cas9) and the RNA-bound(PDB ID- 4ZT0, Cas9+sgRNA) Jiang *et al.* (2015) and RNA-DNA–bound states(PDB ID- 4OO8, Cas9+sgRNA+tDNA) of the enzyme tells about the flexible nature of Cas9 on sgRNa binding and DNA recognition. In Figure 1.5A, Structures of three states of SpCas9 are shown in cartoon representation shown in the free state, pre-target bound state, and PAM-containing duplex DNA target-bound state, respectively. The two lobes of Cas9 and the Arg-helix(bridge helix) are colored differently. In Figure1.5B, the aligned structures of the three Cas9s obtained from PyMol Schrodinger and DeLano (2020*a*) are shown. The struc-

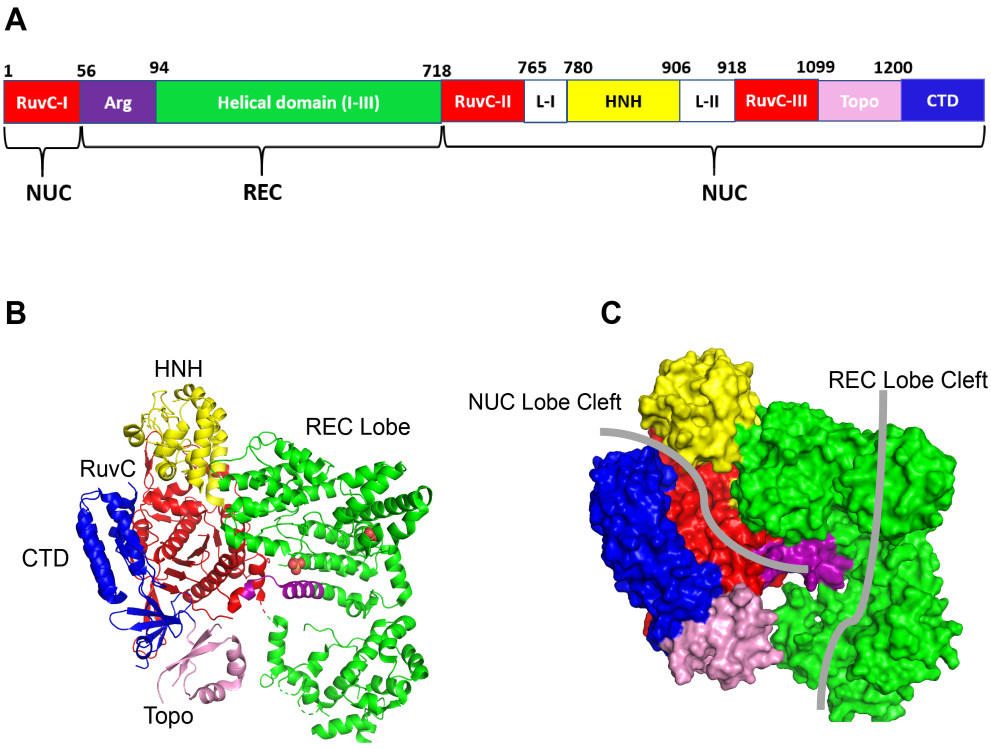


Figure 1.4: A. Domain organization in typical SpCas9 structure. B. Crystal structure of SpCas9 (PDB ID: 4CMP) shown in cartoon representation. Individual Cas9 regions are colored differently. The HNH domain in yellow, REC lobe in green, RuvC domain in red, C-terminal domain in blue, Topo in light pink and bridge-helix is shown in violet color. C. Surface representation of SpCas9 depicting the two nucleic acid binding clefts on the molecular surface.

tural alignment of free and pre-target bound Cas9 has an RMSD of 15.188 Å, and 970 atoms were aligned (Figure 1.5B(i)), while the RMSD of free and ternary complex was very high, 19.758 Å and 939 atoms were aligned (Figure 1.5B(ii)). The RMSD of pre-target bound and ternary complex of Cas9 bound with sgRNA and ssDNA was 6.395 Å with 1051 atoms aligned (Figure 1.5B(iii)). The structural deviations were high, from free to pre-target state as compared to pre-target to ternary complex. From the figures, it is visible that the nuclease lobe in bound states aligned better to the free form of Cas9 as compared to recognition lobes. On binding of sgRNA to apo-Cas9, the REC lobe undergoes a change in conformation, activating the enzyme into a DNA recognition competent conformation. The Hel-III of the REC lobe moves $\approx$ 65 Å toward the HNH domain upon sgRNA binding. But, within the NUC lobe, there were few segments with more significant structural movements. In order to understand the structural movements of each domain or segment within the NUC lobe in different states of Cas9, the NUC lobe was focused (hiding the REC lobe) in Figure 1.5C. Figure 1.5D, clearly shows the structural rearrangement of HNH and RuvC domains on binding to sgRNA and RNA-DNA hetero-duplex. The conformational changes in Cas9 from Cas9+sgRNA state to Cas9+sgRNA+tDNA & PAM sequence state exhibit much smaller conformational changes indicating that the extensive structural rearrangements occur before target DNA binding, reinforcing the notion that guide RNA loading is a crucial regulator of Cas9 enzyme function. Chylinski *et al.* (2014); Chen *et al.* (2019)

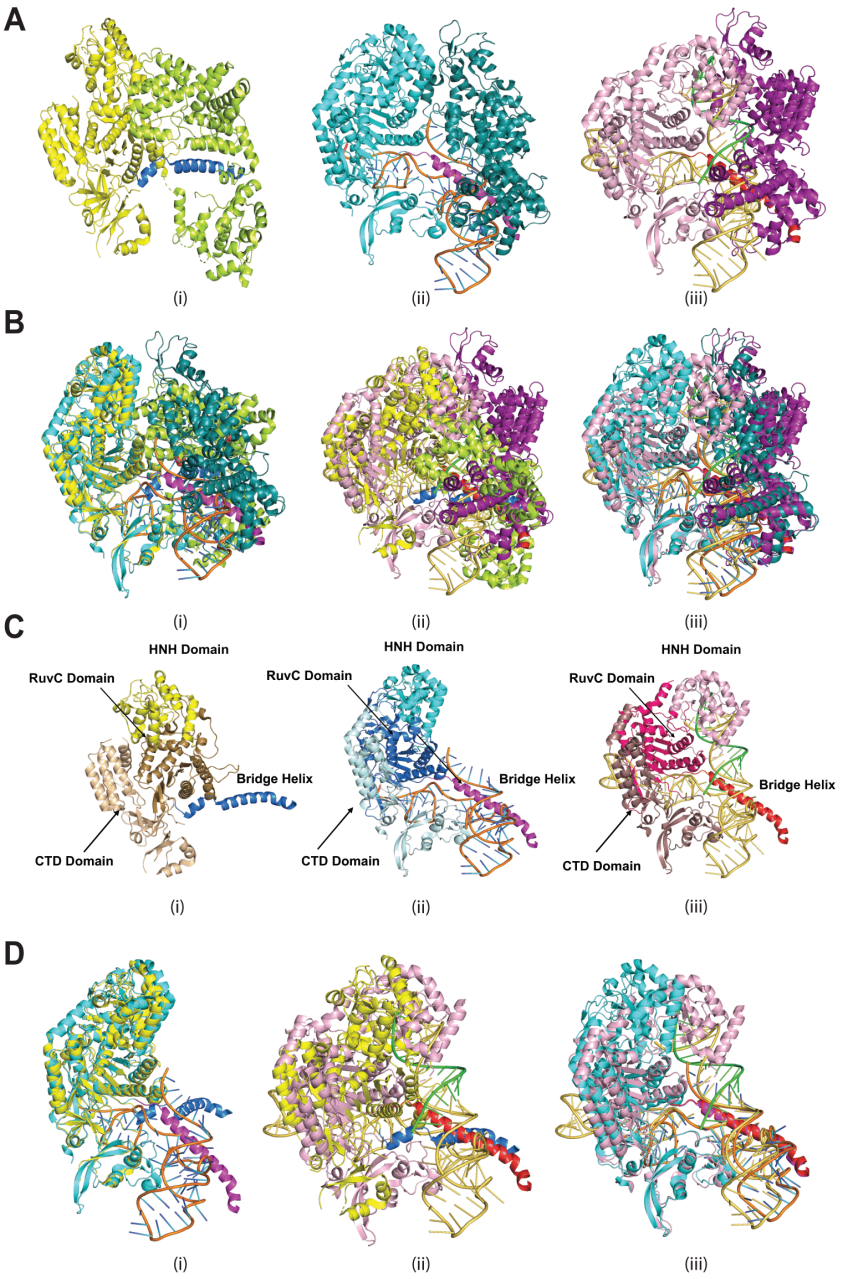


Figure 1.5: Structures of bound and unbound SpCas9s. A(i) Apo-Cas9 (PDB:4CMP) - NUC lobe (yellow) and REC lobe(limon) and Arg helix(blue) (ii) Cas9 binary complex(PDB:4ZT0) - NUC lobe (cyan), REC lobe (teal), and Arg Helix (magenta). (iii) Cas9 ternary complex - Cas9+sgRNA+tDNA (PDB:4OO8) ; NUC lobe (light pink), REC-lobe (deep purple), Arg helix (red), sgRNA (yellow), tDNA (green). B(i) Aligned structure of apo and pre-target bound states of Cas9(4CMP+4ZT0). (ii) Aligned structure of apo and ternary Cas9 (4CMP+4OO8). (iii) Aligned structure of binary and ternary Cas9s (4ZT0+4OO8). C (i)Apo form of Cas9 (4CMP). (ii). Pre-target bound state of Cas9 binary complex (4ZT0). (iii). Single-Stranded DNA bound state of Cas9 ternary complex(4OO8). D. Superposed images of NUC lobe (i) Apo and pre-target bound states of Cas9(4CMP+4ZT0), (ii) Apo and ternary Cas9(4CMP+4OO8). (iii) Binary and Ternary Cas9s(4ZT0+4OO8).

1.4 Applications and Challenges of CRISPR/Cas gene editing

CRISPR/Cas technology has emerged as a transformative tool across various fields of biomedical and agricultural sciences, showcasing its versatility and effectiveness in addressing complex genetic issues. Its applications span from correcting genetic mutations to enhancing agricultural productivity, making it a pivotal innovation in modern science (shown in Figure 1.6).

One of the most significant applications of CRISPR/Cas technology is in cardiovascular diseases (CVDs). Research has demonstrated that CRISPR can correct mutations in the calmodulin 2 (CALM2) gene, which is associated with long QT syndrome (LQTS), a condition that can lead to severe cardiac events. Limpitikul et al. successfully corrected a CALM2 mutation in induced pluripotent stem cell-derived cardiomyocytes, effectively rescuing cardiac function and preventing arrhythmias associated with LQTS Motta *et al.* (2017). Moreover, CRISPR/Cas9 has been utilized to explore the genetic basis of various cardiac conditions, including hypertrophic cardiomyopathy and familial hypercholesterolemia. Studies have shown that targeting specific genes, such as RYR2, can ameliorate symptoms in animal models, indicating potential therapeutic avenues for human applications Roshanravan *et al.* (2022); van Kampen and van Rooij (2019). The ability to create isogenic human cell lines using CRISPR also facilitates drug testing and the development of personalized medicine approaches.

Cancer initiation and spread are mediated by mutations and dysregulation of a variety of genes such as oncogenes, tumor suppressor genes, stem cell-associated genes, chemo-resistant genes, and metabolic genes. Cancer treatment's primary goal is to halt cancer cell growth and development by repairing mutations and restoring dysregulated gene expression. Since its inception, the CRISPR/Cas9 gene-editing method has been widely used in cancer research, with promising results. It can be used to investigate the genesis, progression, and metastasis of tumors. Recently, there has been a growing use of this approach in experimental studies on tumor therapy. The CRISPR/Cas9 system can selectively eliminate genes or rectify mutations to treat cancer. Preliminary tumor treatment investigations have been undertaken in several pertinent disciplines. Molecular-level cancers may be treated using CRISPR/Cas9. Utilising CRISPR/Cas9 technology, personalised and targeted therapies have the potential to revolutionise tumour therapy. Georgiadis et al. recently demonstrated that fratricide-resistant T cells can be generated by removing and replacing the TCR/CD3 and CD7 with lentiviral-mediated production of CARs specific for the CD3 or CD7 Georgiadis *et al.* (2021); Dhainaut *et al.* (2022).

Nervous system disorders have intricate causes, pose considerable challenges in terms of treatment, and significant impact on the quality of life of patients. According to the World Health Organisation (WHO), neurodegenerative disorders are projected to surpass cancer as the primary cause of mortality in humans by the year 2050. Several recent studies have shown that the CRISPR/Cas9 gene-editing technology is highly effective in treating neurological disorders such as Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD), glioblastoma (GBM), epilepsy, and amyotrophic lateral sclerosis (ALS) Safari *et al.* (2020); Yun and Ha

(2020); Carpenter and Lignani (2021). Recently, it has been instrumental in identifying transcriptional regulators involved in diseases such as Glioblastoma multiforme Carpenter and Lignani (2021). Using a CRISPR/Cas9 nano-complex (Cas9-sgRNA augmented with two amphiphilic R7L10 peptides in varying amounts), Park et al. introduced sgRNA carrying Bace1 gene targeting sites into mice models of five familial Alzheimer's disease (5XFAD) and APP knock-in AD. Significant decreases in Bace1 and Aβ were observed, indicating a reduction in the clinical symptoms and cognitive impairment induced by AD. CRISPR/Cas9 has been explored for its ability to introduce mutations that mimic PD in animal models, facilitating the study of disease progression and the evaluation of therapeutic interventions targeting mitochondrial dysfunction and oxidative stress, common features in PD pathology.

CRISPR technology is also making strides in treating ocular diseases as well. For example, it has been employed to restore the reading frame of mutated photoreceptor genes like RPGR by excising specific gene segments. This approach holds promise for reversing vision loss associated with genetic disorders Martinez Velazquez and Ballios (2021). In Leber congenital amaurosis type 10 (LCA10), a mutation in the CEP290 gene leads to gradual vision loss in children. Researchers have used CRISPR to correct this mutation by knocking in the correct gene sequence, effectively restoring CEP290 protein levels and reversing vision loss in animal models Ahmad (2022). CRISPR has also shown promise in treating viral ocular infections, such as herpes simplex virus (HSV) keratitis. This condition, caused by a viral infection of the cornea, can lead to blindness. Researchers have employed CRISPR to disrupt essential viral genes like UL8 and UL29, effectively inhibiting HSV replication and progression. It has also potential in inhibiting viral replication by editing the HBV ORF genes involved in chronic viral infections such as hepatitis B and C. The presence of covalently closed circular DNA (cccDNA) in HBV-infected cells is a major barrier to effective treatment, as current antiviral therapies primarily suppress viral replication without eliminating cccDNA. Recent studies have demonstrated that CRISPR/Cas9 can specifically cleave the HBV genome, leading to a reduction in viral RNA and protein levels, thereby inhibiting replication. For instance, synthetic guide RNAs (gRNAs) targeting various HBV genes have been developed, showing robust antiviral efficacy in cellular models Stone *et al.* (2021); Zhang *et al.* (2023). This application could lead to novel antiviral therapies that directly target the genetic material of the virus. By inducing pluripotent stem cells that secrete anti-cytokine drugs in response to inflammation, this technology offers a novel therapeutic strategy in treating autoimmune disorders such as rheumatoid arthritis that could significantly alter the management of autoimmune diseases. This approach could significantly alter the management of autoimmune diseases by providing a targeted and responsive treatment mechanism, reducing the need for systemic immunosuppression. CRISPR's precision has also transformed molecular diagnostics and genome-wide screening, allowing for the identification of potential drug targets and genetic markers with unprecedented accuracy. The development of CRISPR-based diagnostic tools, such as the SHERLOCK and DETECTR assays, has demonstrated the ability to detect specific nucleic acid sequences with high sensitivity and specificity, making it possible to identify pathogens and genetic mutations rapidly and affordably. These methods are not only effective in clinical settings but also show promise for point-of-care applications, allowing for timely diagnosis in resource-

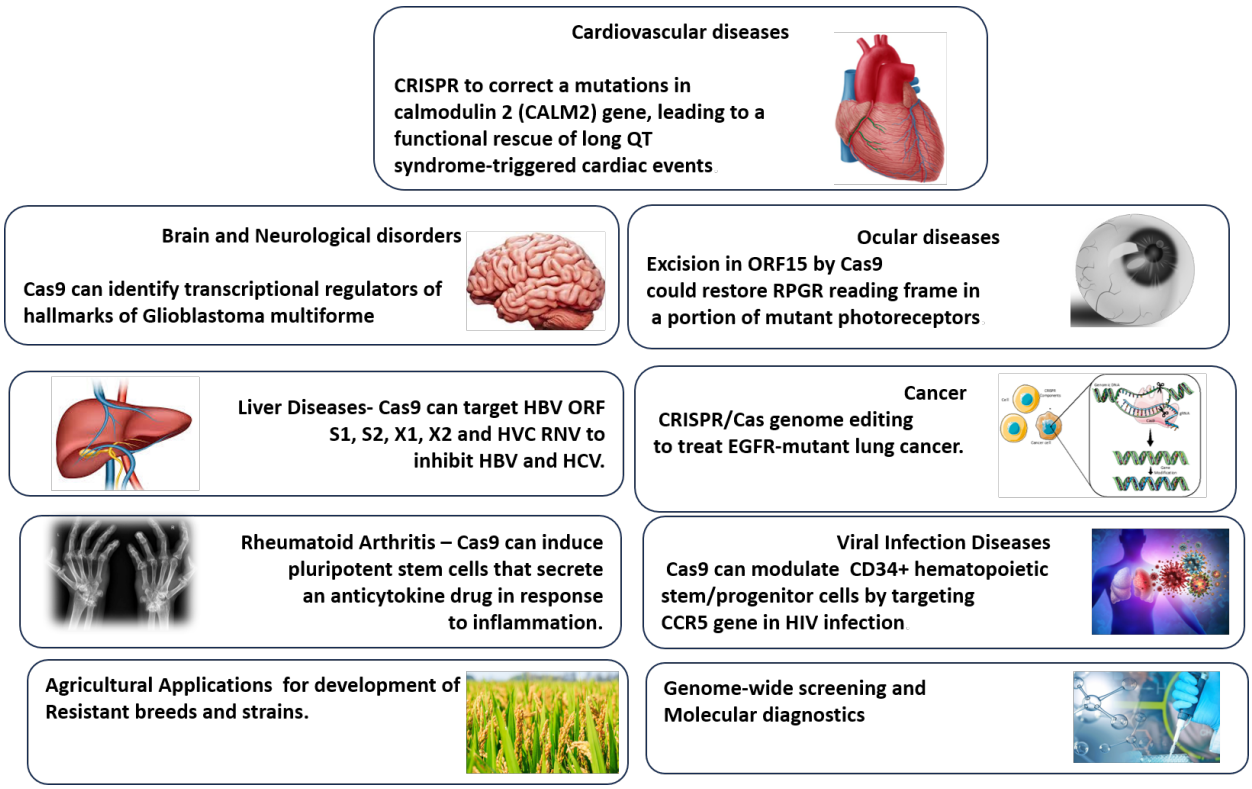


Figure 1.6: CRISPR/Cas9 system has been utilized to create disease-based models for various significant human diseases such as viral diseases, neurological diseases, cancer, ocular disease, blood diseases, cardiovascular diseases and disorders, and other complex genetic human diseases.

limited environments Ghouneimy *et al.* (2022). Beyond human health, CRISPR/Cas technology is revolutionizing agriculture by enabling the development of resistant plant breeds and strains. This genetic editing capability enhances crop resilience to diseases and environmental stressors, thereby improving food security. CRISPR has been effectively used by researchers to develop crops resistant to particular infections, which lowers the demand for chemical pesticides and supports sustainable farming methods. CRISPR technology increases crop resilience, which boosts agricultural output and promotes environmental conservation efforts by reducing chemical inputs and the ecological effects Liu *et al.* (2021); Zhu *et al.* (2020)

Despite the above explanation indicating that CRISPR/Cas9 shows promise, this editing system retains some constraints and challenges that pose concerns for its application in clinical trials, mostly due to its recent discovery and use in humans. Several significant issues with the CRISPR/Cas9 technology, including as immunogenicity, off-targeting, polymorphism, delivery technique, and ethics, have been identified. The following list of strategies has been devised to address these limitations (Figure 1.7).

Major concerns related to CRISPR/Cas9 Gene Editing

Immunogenicity

An inherent issue associated with CRISPR/Cas9 is its possible immunogenicity. Due to its bacterial origin, the Cas9 protein can provoke an immunological response in humans, resulting in diminished effectiveness of the treatment and possible safety concerns Mehta and Merkel (2020*a*). Likewise, previous studies have documented the presence of an existing immune response to the identical Cas9 protein in Staphylococcus aureus Colque-Navarro *et al.* (2010). Indeed, both Staphylococcus aureus and Streptococcus pyogenes, the primary sources of the Cas9 proteins SaCas9 and SpCas9, have been infecting humans for an extended period of timeCharlesworth *et al.* (2019*b*). Hence, the human immune system identifies these proteins as exogenous and initiates an immunological reaction against them upon administration, resulting in rapid breakdown of the Cas9 protein, so inhibiting their ability to carry out the gene-editing functioing Ewaisha and Anderson (2023); Mehta and Merkel (2020*a*).

Delivery Techniques

One additional problem that must be addressed is selecting a suitable, secure, and precise delivery method to transport the CRISPR system into the tumor location, particularly in vivo, and specifically targeting the correct sequence within the nucleus. Delivery of CRISPR/Cas9 technologies incorporates several methods, including viral, physical, and extracellular vesicle-based system delivery mechanisms Yip (2020). Moreover, each approach is used for a certain objective and inherently has its constraints. Therefore, selecting the appropriate vector are packing, distribution, and accurately targeting the intended location becomes crucial Mout *et al.* (2017). Current delivery methods, such includes viral vectors and non-viral vectors. Adeno-associated viruses (AAV) and lentiviruses are popular viral vectors in gene therapy due to their safety and low immunogenicity. AAVs are favored for their non-pathogenic nature and minimal immune response in humans, but their small packaging capacity (around 4.7 kbp) limits their ability to deliver larger gene editing systems like CRISPR in a single vector. This often requires the use of multiple AAVs for complete delivery, complicating the process. On the other hand, lentiviruses can carry larger genetic material and integrate into the host genome, though this raises concerns about insertional mutagenesis.Non-viral delivery systems are being explored, but they often face limitations in terms of efficiency and specificity Yip (2020); van Haasteren *et al.* (2020). Non-viral delivery vectors, such as lipid nanoparticles and inorganic nanoparticles, offer an alternative to viral vectors in gene therapy. It allows frequent administration with reduced immunogenicity, better protection from nucleases, and improved targeting. Also, non-viral vectors can carry larger genetic material without integrating into the genome. However, challenges remain in achieving efficient delivery and high editing efficiency, particularly in applications like cancer therapy Chen *et al.* (2020*a*). By utilising recombinant CRISPR-Cpf1 Ribonucleoprotein (CRISPR-Cpf1-RNP) and smaller Cas9, the risk of off-targeting can be reduced.

Off-target Activity

CRISPR/Cas9 can sometimes introduce unintended mutations at sites other than the intended target, known as off-target effects. These unintended edits can lead to adverse consequences, including the activation of oncogenes or the disruption of essential genes, raising concerns about the long-term safety of CRISPR therapies In recent years, many bioinformatics tools have been developed to help predict and reduce off-target modifications. These should be further improved to enable researchers to use them effectively in the development of new therapies. It is further discussed in detail in the next section Lessard *et al.* (2017).

Ethical Considerations

The use of CRISPR technology raises ethical concerns, particularly regarding germline editing, which can have lasting implications for future generations. The potential for "designer babies" and unintended consequences in the human gene pool has sparked significant debate among scientists, ethicists, and policymakers. The application of CRISPR/Cas9-based gene editing for human germline alteration has generated apprehensions over the endangerment of human life and dignity, as well as the possibility of genocide. The moral comcepts in biomedicine center on evaluating the possible hazards and advantages of a novel technology such as CRISPR. There exist three main concerns pertaining to CRISPR genome editing. Firstly, CRISPR has inherent technical constraints, including such as low editing accuracy, off-target effects, and incomplete edits. Second, there are concerns about the long-term effects on edited species and whether the changes will be passed down through generations in ways that we can't fully predict. Lastly, even if the gene is edited as planned, we still don't fully understand how genes and biological traits are linked, making it hard to predict the real impact of these changes. Because of these uncertainties, making moral decisions about the use of CRISPR is complex and challenging.

Gene Polymorphism in Cancer

Cancer polymorphism refers to the presence of diverse genetic mutations within tumor cells, leading to a heterogeneous population that can complicate treatment. CRISPR/Cas9 technology faces challenges in addressing this complexity, as cancers often harbor multiple mutations across various genes, making it difficult to design a single editing strategy that effectively targets all relevant mutations. Additionally, off-target effects can introduce unintended mutations, potentially exacerbating tumorigenesis Gonatopoulos-Pournatzis *et al.* (2020); Aregger *et al.* (2021). To mitigate these issues, researchers can employ multiplexed CRISPR approaches that simultaneously target multiple genes, improving the chances of effectively addressing the tumor's genetic landscape. Enhanced delivery methods, such as nanoparticle systems or viral vectors, can facilitate the introduction of multiple guide RNAs and Cas9 into cancer cells. Furthermore, focusing on the tumor microenvironment and utilizing advanced techniques like base editing can help reduce off-target effects and improve the safety and efficacy of CRISPR-based therapies.

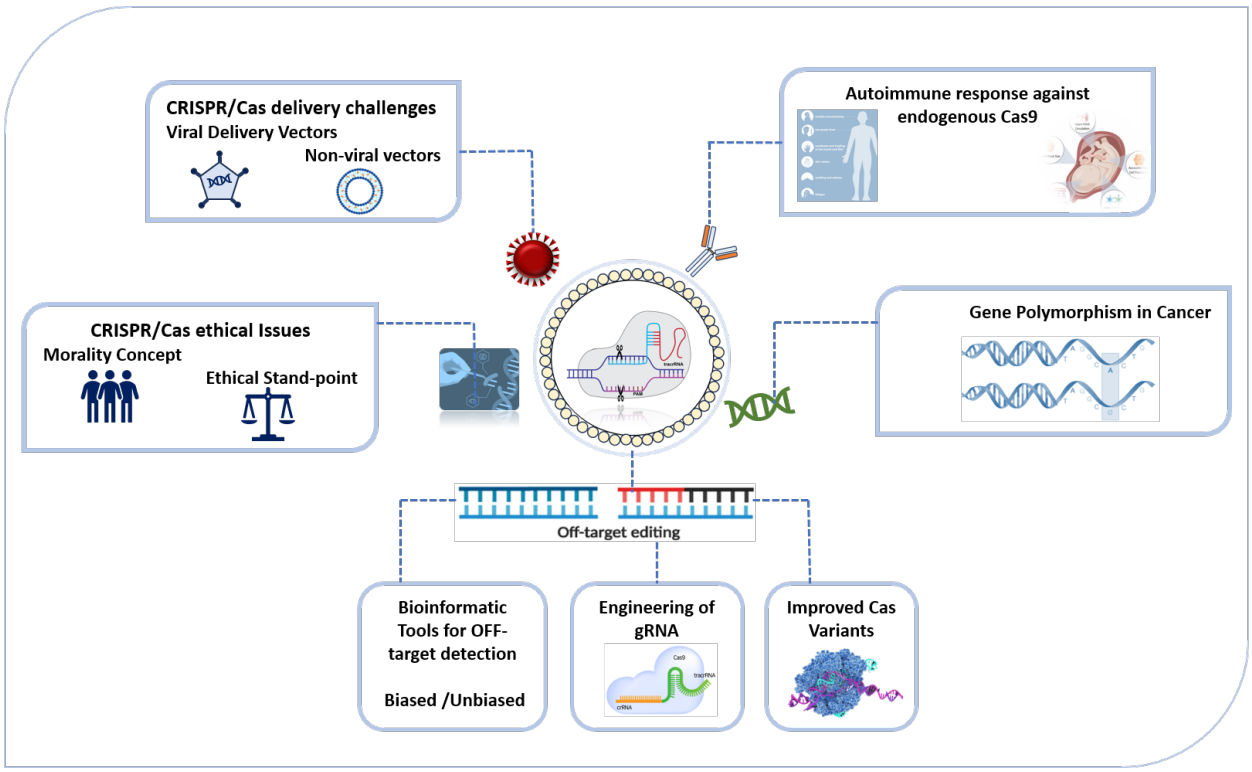


Figure 1.7: CRISPR/Cas9 system has been utilized to create disease-based models for various significant human diseases such as viral diseases, neurological diseases, cancer, ocular disease, blood diseases, cardiovascular diseases and disorders, and other complex genetic human diseases.

1.5 Prediction and Mitigation of Off-target activity

CRISPR/Cas-based genome editing technologies have been shown to be indispensable for scientific progress because of their wide range of applications in both basic research (health, agriculture, etc) and the development of biotherapies for complicated illnesses like cancer. Despite the high efficiency and user-friendliness of CRISPR/Cas9, its applications are limited by various factors which includes low efficiency of homology-directed repair, dependency of SpCas9 on specific PAM sequence (NGG) limits the potential target sites, the large size of Cas9 hinders its packaging and delivery to the target site by AAV vectors Ren *et al.* (2017), and immunogenic behaviour of Cas9 Charlesworth *et al.* (2019*a*),etc. The off-target activity of Cas9, major concern, involves insertion of DNA mismatches in PAM-distal end of the sgRNA sequence which can lead to chromosomal rearrangement and mutations Mali *et al.* (2013); Cong *et al.* (2013*a*). Several groups have talked about the propensity of gRNA-Cas9 complexes to cleave off-target DNA sequences. Pattanayak and his group Pattanayak *et al.* (2013*a*) took eight gRNA-Cas9 complexes, ten potential off-target DNA sequences and used in-vitro selection and high-throughput sequencing for this study. It was observed that target specificity of gRNA-Cas9 is due the seed sequence which is 10-12 bp proximal to the PAM motif in the 3'end of crRNA. From other studies, it was revealed that the off-targets DNA cleaved by Cas9 can accommodate upto 5 mismatches in their seed-sequence. Mismatches in the sgRNA seed sequence can drastically impede target DNA binding and cleavage,

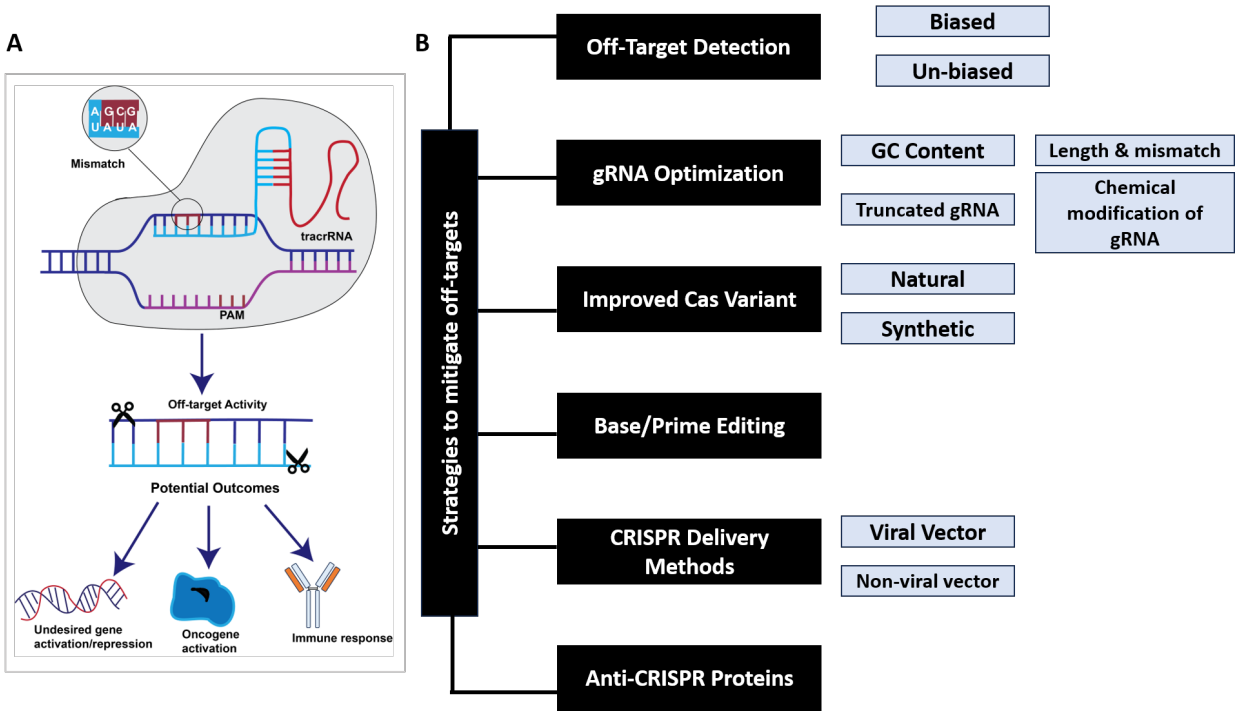


Figure 1.8: Off-target effects by CRISPR/Cas9 and strategies to mitigate it.

whereas close homology in the seed region commonly leads to off-target binding events even when there are many mismatches elsewhere Pattanayak *et al.* (2013*a*). The configurations and architectures of gRNAs inside the system generally exert an influence on off-target behavior. In addition to potentially throwing doubt on scientific findings on the activity of genes, off-target genome editing confuses and limits the therapeutic potential of CRISPR-Cas9. As illustrated in Figure 1.8, this effect can result in negative outcomes such as unwanted DNA damage, immunological response, and cytotoxicity.

1.5.1 Strategies to mitigate off-target effects of CRISPR/Cas9

Off-target Detection Methods

Biased Method In-silico predicitve models can predict and quantify off-target effects which are categorized as (a) Alignment-based method and (b) Scoring-based method. Alignment based models rely on matching of guide RNA (gRNA) sequences with reference genome sequences to detect possible off-target effects. These models encompass a range of tools, including Bowtie, BWN, CasOT, Cas-OFFinder, FlashFry, and CrisFlash. Bowtie and BWN algorithms detect off-targets with a maximum of one mismatch, but CCTop utilizes Bowtie to detect off-target activity, allowing for a maximum of five mismatches. CRISPOR use Bayesian Weighted Analysis (BWA) to model all possible off-target locations. Cas-OFFinder model is capable of detecting off-targets with single base pair indels by accommodating different PAM sequences, mismatches, and gRNA lengths Bae *et al.* (2014); Xiaorae4 *et al.* (2014). FlashFry rapidly detects off-targets and offers GC con-

tent and on/off-target ratings, whereas CrisFlash employs a tree-based machine learning approach to improve the accuracy of gRNA design Jacquin *et al.* (2019). Cas-OFFinder is renowned for its extensive off-target detection capabilities, while FlashFry is esteemed for its rapidity and effectiveness McKenna and Shendure (2018). Scoring-based techniques for predicting off-target effects in CRISPR/Cas9 gene editing includes MIT, CFD, CRISTA, Elevation, and DeepCRISPR methods. The MIT (Hsu-Zhang) score, which is included into programs such as CHOPCHOP and CRISPOR, assesses potential off-targets by analyzing the locations of mismatches in the 20 base pair target sequence Labun *et al.* (2016). CRISTA is a rigorous off-target evaluation tool that takes into account variables such as GC content, RNA structure, and epigenetics to find cleavage efficiency Abadi *et al.* (2017). It outperforms CFD, CCTop, and MIT. An advanced two-layer regression model called Elevation integrates sequence and DNA accessibility/epigenetic data to accurately forecast individual and combined off-target scores for the human exome. It surpasses the performance of MIT, CFD, and CCTop models, but its accuracy is restricted to a single genome. The DeepCRISPR computational biology program accurately predicts both on-target and off-target cleavage sites, taking into consideration epigenetic characteristics that impact effectiveness Chuai *et al.* (2018). Elevation method is widely regarded as the most sophisticated and effective approach for achieving thorough off-target identification Listgarten *et al.* (2018).

Unbiased Method Unbiased methods can detect off-target sites in living cells on the whole genomic level. These methods are categorized as in vitro detection or in vivo genome-wide assays or cell-based genome-wide assays. In vitro genome-wide assays for detecting off-target effects in CRISPR/Cas9 include Digenome-seq, SITE-seq, and CIRCLE-seq. Digenome-seq is a robust method for genome-wide off-target detection with a sensitivity of 0.1% or lower, but requirement of high read depth, making it less suitable for screening large numbers of gRNAs Kim *et al.* (2015). SITE-seq uses Cas9-sgRNA ribonucleoproteins (RNPs) to cleave purified genomic DNA in a cell-free system, allowing for the tagging and sequencing of both on- and off-target cleavage fragments, while requiring less sequencing depth, hence, improving upon Digenome-seq. CIRCLE-seq follows a similar concept but involves circularizing DNA to eliminate background noise before cleavage, enhancing sensitivity and allowing the detection of off-target sites influenced by cell-type-specific SNPs Cameron *et al.* (2017*a*). However, it requires a large quantity of genomic DNA, which can be a limiting factor. Both SITE-seq and CIRCLE-seq can detect more cleavage sites than in vivo methods due to the absence of chromatin structure and associated proteins Crosetto *et al.* (2013). While CIRCLE-seq can detect up to 94% of off-target sites, CRISLE-seq is effective in specific cells labeled with GUIDE-seq tags. Overall, these in vitro methods provide valuable insights into off-target effects, although they may not fully replicate the complexities of in vivo environments. In vitro and cell-based cleavage conditions differ significantly, leading to the realization that many off-target effects identified in vitro may not occur in vivo due to cellular repair mechanisms. This necessitated the development of in vivo or cell-based genome-wide assays to accurately detect off-target effects. Commonly used in vivo methods include integrase-defective lentiviral vectors (IDLVs), chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), breaks labeling enriched on streptavidin and next-generation sequencing (BLESS), genome-wide unbiased identification of double-strand breaks

(DSBs) enabled by sequencing (GUIDE-seq), the discovery of in situ Cas off-targets verified by sequencing (DISCOVER-seq), etc. IDLVs were among the first tools for identifying genome-wide off-target sites, integrating with DSBs and allowing for PCR analysis of distribution. However, they have limitations, including a low detection efficiency of about 1% and an inability to analyze multiple off-target sites Wang *et al.* (2015). ChIP-seq, which uses a dead Cas9 (dCas9) to bind DNA without inducing DSBs, can identify off-target sites but may alter specificity. BLESS is a newer method that offers greater sensitivity and versatility for mapping DSBs but requires a reference genome and only detects off-target effects during the labeling period Park (2009). GUIDE-seq, which incorporates oligodeoxynucleotides into DSBs, is sensitive but requires high sequencing read depth to minimize background noise Joung and Tsai (2015). LAM-HTGTS is another robust method that detects chromosomal translocations due to DSBs, but it only identifies specific translocations, which may limit its utility Wienert *et al.* (2020). DISCOVER-seq is a sensitive assay that utilizes the MRE11 repair factor to identify off-target sites without introducing exogenous tags, making it applicable to human and mouse cell lines Wienert *et al.* (2020).

gRNA Optimization

GC Content Structural analysis has revealed that gRNAs with GC contents between 40% and 60% exhibit increased on-target activity Doench *et al.* (2014); Wang *et al.* (2014). Higher GC content stabilizes the DNA-RNA duplex and destabilizes off-target binding. Additionally, the positioning of purine residues at the end of four nucleotides and the preference for guanine at position 20 and cytosine at position 16 can enhance editing efficiency Wong *et al.* (2015). Consistent with these findings, a positive correlation has been observed between the GC percentage in the PAM-proximal region of sgRNAs and their mutagenesis efficiency Ren *et al.* (2014).

Length and Base-pair mismatch gRNAs with lengths up to 17 nucleotides have shown higher genome-editing efficiency, while lengths of 18-20 base pairs (bp) have resulted in lower efficiency . Interestingly, 20 bp gRNAs have not been observed to be associated with any unwanted mutatioFu *et al.* (2014*a*). To further reduce off-target effects, strategies such as partial DNA replacement and the use of chimeric guide RNAs have been explored in primary human cells Sugano *et al.* (2018). Guidelines have been established to minimize potential off-target effects, including avoiding target sequences with more than three mismatches within 7-10 bp of the PAM and preventing sgRNA bulges within 12 bp of the PAM Yin *et al.* (2018); Lin *et al.* (2014).

Truncated gRNA In human cells, Cas9s with truncated sgRNAs (17- or 18-nt guide sequences) showed significantly reduced off-target activity while maintaining on-target genome editing efficiency Slaymaker *et al.* (2016*a*); Fu *et al.* (2014*a*). However, upon combining, a gRNA with mismatched G at the 5' and the truncated sgRNA with a high-fidelity Cas9 variant, SpCas9-HF1, an impaired on-target activity was observed Kleinstiver *et al.* (2016*b*). In support of this observation, a similar study performed by Zhang and group, revealed that for perfect on-target editing activity of eSpCas9 and SpCas9-HF1, single guide RNAs with accurate, completely matched 20-nucleotide guide sequences are required Zhang *et al.* (2017). The on-target activity was also

reduced by using a 3' end-truncated sgRNA or a 5' end-added sgRNA (-GG) Feng *et al.* (2021). Thereby, limiting its wide-spread application in reducing the off-target affects in SpCas9 mediated gene-editing.

Chemical Modification of gRNA Interestingly, chemical alterations in the 20-nucleotide (nt) guide sequence at the 5' end of a gRNA, which mediates Cas9 specificity in identifying and hybridizing complementary protospacers, can be exploited to fine-tune the specificity of CRISPR–Cas systems. Ryan et al. demonstrated that CRISPR–Cas9 indel rates and homology-directed repair (HDR) editing events can be significantly increased by chemical modifications in the first and last three-terminal nucleotides of sgRNAs Ryan *et al.* (2018). Incorporating phosphonoacetate linkages and 2'-O-methyl (known as 2'-O-methyl-3'-phosphonoacetate) at specific positions of the guide sequences improved Cas9 specificity and maintained comparable on-target activity of unmodified sequences Chen *et al.* (2021). In a work by Donohoue et al. in 2021, it was shown that 2'-deoxynucleotide (dnt) positioning affects guide activity and specificity and can be used for engineering CRISPR-hybrid-RNA-DNA with reduced off-target effects while preserving on-target editing activity Donohoue *et al.* (2021). These findings suggest a novel way for improving specificity in gRNAs by chemically altering the guide sequence.

Improved Cas Variant

Naturally Occuring Cas9 Orthologs The structural alignment of Cas9 nuclease domains with other DNA-bound nucleases disclosed that RuvC like domain shares a structural similarity with the integrase superfamily members. These retroviral integrase proteins are known for their RNase H fold, indicating that the RuvC domain might be using a two-metal ion catalytic mechanism for cleaving non-target DNA strands. The HNH-domain shares structural similarity to HNH-endonucleases, exhibiting a signature $\beta\beta\alpha$ fold suggesting that it employs one- metal ion mechanism for cleavage of the target DNA strand Jinek *et al.* (2014*a*); Pattanayak *et al.* (2013*a*). A typical feature of the one/two metal-ion dependent nucleases is that it has two conserved residues, histidine in HNH-domain (His840) and aspartate residue in the RuvC domain (Asp10). Mutation of any of these domains converts the Cas9 into nickase (Cas9n), which forms a single-strand break and mutation of both these residues renders it in-active or dead(dCas9). It can recognize the DNA, but its ability to cleave it is withdrawn Jinek *et al.* (2014*a*). A single Cas9 protein can exhibit a unique activity at multiple sites, but towards the exhibition of different activity at multiple sites, one requires multiple Cas9–gRNA orthologs Braff *et al.* (2016); Gasiunas *et al.* (2020). Additionally, the most widely Streptococcus pyogenes Cas9 (SpCas9) has limitations, particularly in generating off-target mutations. To address these issues, researchers have talked about the different orthologs of Cas9 including SaCas9, St1Cas9, St3Cas9, NmCas9, FnCas9, and CjCas9, each offering distinct advantages like higher efficiency and specificity , smaller size and wider PAM-activity, etc Fuchsbauer *et al.* (2019); Hirano *et al.* (2019); Nishimasu *et al.* (2015); Yamada *et al.* (2017). SaCas9, derived from Staphylococcus aureus, is smaller than SpCas9, making it suitable for packaging in adeno-associated virus (AAV) vectors. It recognizes a longer protospacer adja-

cent motif (PAM) sequence (5'-NNGRRT-3'), potentially reducing off-target effects by increasing the frequency of PAM recognition in the genome. St1Cas9 and St3Cas9, derived from Streptococcus thermophilus, exhibit higher on-target activity compared to wild-type SpCas9, recognizing PAM sequences of 5'-NNAGAAW-3' and 5'-NGGNG-3', respectively. NmCas9 from Neisseria meningitidis demonstrates high specificity with its PAM site (5'-NNNNGHTT-3') but has lower on-target activity than SpCas9. FnCas9, derived from Francisella novicida, is unique in its ability to target both DNA and RNA, although its on-target activity is lower than that of SpCas9. Among these orthologs, CjCas9 has emerged as a promising candidate due to its small size and high on-target activity, making it easily transferable via AAV delivery methods. This ortholog shows great potential for therapeutic gene editing applications.

Engineered Cas variants Off-target activity was also decreased by mutations in amino acid residues involved in stabilising the R-loop structure. The R-loop is a nucleic acid structure made up of RNA strand base-paired to one strand of DNA and the other strand of DNA remaining unpaired. In compliance with this concept, four Cas9 variants with high fidelity and enhanced specificity have been engineered, namely (enhanced specificity Cas9 (eSpCas9) and Cas9-high fidelity (SpCas9-HF)) Kim *et al.* (2017); Anderson *et al.* (2018) which have higher specificities than wt-Cas9, HypaCas9, and evoCas9 Chen *et al.* (2017*b*); Casini *et al.* (2018*a*). The evoCas9, in particular, has a 79-fold higher fidelity than wt-Cas9. Cas9 nickases formed after inactivating one of the catalytic residues of Cas9 were also found to have reduced off-target activity and facilitated overhang-based cloning Ran *et al.* (2013*b*,*a*). Cas9 variants with better specificity were designed using the structural information obtained from all previous MD investigations. Engineered Cas9 variants, SpCas9-HF1 and eSpCas9(1.1), had shown decreased cleavage activity when delivered using the therapeutically active ribonucleoprotein delivery method. A high-fidelity Cas9 variant, Hifi Cas9 (R691A) when employed as an RNP complex, decreases global off-target effects while maintaining high on-target activity. Vakulskas *et al.* (2018*a*); DeWitt *et al.* (2016). In a work published by Jonathan et al, tagmentation-based tag integration site sequencing (TTISS), an efficient, scalable method for analyzing double-strand breaks (DSBs) was used for screening the specificity and activity of Cas9 variants. evoCas9 was found to be the most specific variant with the least on-target activity suggesting a trade-off between Cas9 activity and specificity Casini *et al.* (2018*a*). A new variant, LZ3 Cas9 (N690C T769I G915M N980K) exhibits high activity, increased specificity, and a differential +1 insertion profile as compared to WT SpCas9 Schmid-Burgk *et al.* (2020). A recent study led by Jack Bravo and his team provided a new Cas variant, the SuperFi-Cas9 which is 4000–times less likely to cut off-target sites on DNA without compromising its cleavage efficiency. They implemented kinetics-guided Cryo-EM to determine the structure of Cas9 at different stages of mismatch cleavage. SuperFi-Cas9 variant was developed by performing mutagenesis of mismatch-stabilizing residues in the RuvC domain loop. This variant reduced the off-target DNA cleavage but maintained the rapid on-target DNA cleavage Bravo *et al.* (2022*b*). Over a dozen possible mutation sites in Cas9's NUC lobe have been proposed by Zuo and Liu for generating Cas9 variants with improved specificity Zuo and Liu (2017); Yoon *et al.* (2019). Studying the dynamics of Cas9 variants using MD simulation have also aided in understanding off-target effects. Palermo and colleagues published a paper in 2017 Palermo *et al.* (2017) that used GaMD simu-

lations to decode the mechanism of off-target binding at the molecular level. They characterised the molecular determinants of off-target binding at the atomic level using these MD simulations, providing crucial insights into the mechanism of off-target effects in CRISPR/Cas9. Based on this observation, Ray et al. validated this model by a combination of MD simulations (1.5μs, PDB code: 4UN3 with ntDNA) and electron paramagnetic (EPR) measurements of spin-labeled CRISPR/Cas9 ternary complexes, with nitroxide spin labels on the DNA backbone. Neutralizing these residues at the interface could weaken the binding of ntDNA strand to the protein leading to rehydridization between the tDNA and ntDNA strands, causing "R-loop collapse" Ray and Di Felice (2020). A classical all-atom MD simulations of ternary complexes of wild-type SpCas9 and three of its mutants: Lys855Ala, His982Ala and the combination Lys855Ala & His982Ala was executed for over a cumulative time scale of 10s. This study's conclusion emphasises the important structural effect of mutation on Cas9 specificity. In complexes with Cas9 mutants, it was discovered that the "unwound" portion of the non-target DNA strand has more versatility and seeks to shift away from the HNH/RuvC interface, where it is otherwise stabilised by electrostatic interactions in the wild-type complex.

Base/Prime Editors

DNA base editors allow the insertion of point mutations in the DNA without generating DSBs. It is a highly promising approach for the treatment of genetic diseases caused by point mutations (β-Hemoglobinopathies) with relatively more efficacy than homology-directed repair in quiescent cells, such as hematopoietic stem cells. The cytidine base editors (CBEs) allow C>T conversions and adenine base editors (ABEs) allow A>G conversions allowing higher efficiency, specificity, accessibility to previously inaccessible genetic loci, and multiplexing while maintaining a low rate of Indels Antoniou *et al.* (2021). BEs' low frequency of DSB production prohibits the generation of large chromosomal rearrangements from occurring, preserving DNA integrity. Furthermore, base editing allows for multiplex editing of two or more loci, which has shown to be particularly promising in the treatment of blood diseases Rees *et al.* (2019). Since the current pool of BEs only allows for A>G, C>T, and C>G conversions, more enzymes will be needed to generate all of the different types of conversions, also to implement base editing therapeutic methods in the clinical world, the delivery of BEs as mRNA or RNP in clinically relevant cells must be further enhanced Antoniou *et al.* (2021). In a recent Cryo-EM work by Lapinaite et al., a 3.2-angstrom resolution structure of ABE8e bound to DNA was presented which reveals how ABE8e edits DNA bases and could inspire future base-editor design Lapinaite *et al.* (2020). Prime-editing is a ground-breaking method that can produce targeted insertions or deletions, as well as direct install precise transition and transversion mutations at specific genomic loci, making it a versatile tool. An advantage of prime-editing over base-editing could be the presence of bystander mutations especially when multiple Cs or As is present in the editing activity window and less PAM restrictions than other methods such as HDR, NHEJ, or base editing. More advancements are expected in high-throughput and precise gene editing, editing mitochondria or chloroplast genomes, editing plant genomes regardless of species, and without transgenic insertion Nadakuduti and Enciso-Rodríguez

(2021).

CRISPR Delivery Methods

AAVs offer advantages such low immunogenicity and lack of connection with human diseases, but have disadvantages like small packaging capacity and persistent Cas9 expression leading to greater off-target effects Naso *et al.* (2017). Adeno Virusess exhibit little integration capability, but stimulate immunological responses Hanlon *et al.* (2019). Non-viral delivery methods like electroporation, lipid nanoparticles, and cell-derived nanovesicles can transport CRISPR components as ribonucleoprotein (RNP) complexes, minimizing off-target effects by narrowing the editing time window. Lino *et al.* (2018) Polyethyleneimine (PEI) magnetic nanoparticles offer a highly promising non-viral approach that is both non-toxic and well-suited for reducing off-target alterations Hui (2008). Although viral vectors have excellent in vivo transduction efficiency, there are still significant concerns about their immunogenicity, integration, and off-target effects. Therefore, continuous study is required to enhance their safety and effectiveness for clinical use.

Anti-CRISPR Proteins

Anti-CRISPRs (Acrs) are proteins made by viruses and other mobile genetic elements that can effectively block CRISPR/Cas systems. Acrs belong to a class of phage genes called ancillary genes. These genes do not have any strict requirements for the phage's survival, but they are conserved across phage genomes, indicating that they may have played a role in phage evolution in some situations Liu *et al.* (2020). Anti-CRISPRs prevent the DNA binding of CRISPR/Cas proteins, such as Cas9 and Cas12a, to target DNA for genome editing, allowing Cas9 or Cas12a to be turned off in the cellular environment ensuring the accuracy and safety of CRISPR-based therapies Rauch *et al.* (2017); Wu *et al.* (2020); Watters *et al.* (2018); Kim *et al.* (2019); Marino *et al.* (2019). Furthermore, such applications may reduce safety issues while avoiding DNA sequence modifications. Acrs have their own set of problems, such as being difficult to design, taking a long time for identification and optimization, and being prone to interacting with other targets in mammalian cells. The discovery and functional screening of novel Acrs, as well as the understanding of their inhibitory mechanisms, can help improve the use and regulation of CRISPR-gene editing Maji *et al.* (2019).

1.6 FnCas9: The Future of Precise and Safe Genome Editing

Differential activity of Cas9 orthologs could be related to their sequence composition or structural variation. A better understanding about the reason behind the difference in activity of the orthologs can be obtained from their domain organization and structural analysis of their domains. All the Cas9 proteins with their structures available in UniProt (Consortium (2016))

were taken into consideration. There was a total of eight Cas9 proteins with PDB IDs - 4CMP, 5AXW, 5B2Q, 4OGC, 6KC8, 6RJ9, 6JOO, and 5X2H for organisms S. pyogenes (SpCas9), S. aureus (SaCas9), F.novicida (FnCas9), A.naeslundii (AnaCas9), N.meningitidis (NeimCas9), S.thermophilus (St1Cas9), C.diphtheriae (CdCas9) and C.jejuni (CjCas9), respectively. A literature survey of all the relevant papers was done to retrieve information about the length of each domain and its organization Fuchsbauer *et al.* (2019); Hirano *et al.* (2019); Nishimasu *et al.* (2015); Yamada *et al.* (2017). A horizontally stacked bar plot representing the architecture and length of each domain in Cas9 ortholog is shown in Figure 1.9A. The length of all the domains except the REC lobe is more or less similar. The binding of the target DNA strand takes place in the REC lobe, while the NUC lobe is majorly involved in the binding of sgRNA and cleavage of tDNA. The alpha-helical lobe has a crucial role in sensing the nucleic acid binding, regulating the conformational transition of HNH domains, locking it at the cleavage site, and thus catalyzing the cleavage process Palermo *et al.* (2018). This varying length of REC lobe among various orthologs could have an effect on the efficiency of the binding and cleavage process, which might be a reason behind the different rates of activity among different orthologs. The FnCas9 protein known for its lower off-target activity exhibited largest REC lobe suggesting a distinct mechanism of action.

Reports have suggested that the electrostatic distribution of Cas9 has a significant contribution to its activity. The change in electrostatics of the cavity, i.e., formation of a positively charged RNA binding cavity from free to R-exp state (intermediate state) and pre-RNA bound the state of Cas9, was observed by the Palermo group Palermo *et al.* (2017). This finding indicated that RNA recruitment initiates with exposure to the R-rich region, followed by the formation of a positively charged cavity that mediates the accommodation of the nucleic acids. The charge distribution also helps in the regulation of specificity of the Cas9 enzyme. Slaymaker et al. Tangprasertchai *et al.* (2017) designed and engineered Cas9 variants with altered charge distribution and found that neutralization of positive charges in the non-target-groove can lead to decrease in off-target indel formation while preserving on-target activity. Since DNA/RNA are highly charged polymers of nucleic acids, any alteration in charge distribution will adversely impacton electrostatic interaction between the Cas9-sgRNA/Cas9-sgRNA-tDNA complex resulting in destabilization of the complex Chen (2019). In a paper published in 2019, by Debojyoti et al., FnCas9 was found to have very high electrostatic potential as compared to its ortholog SpCas9 which could be one of the reasons leading to its stringent interrogative behaviour and less tolerance to mismatches, hence reduced off-target effects Acharya *et al.* (2019*a*). The previous studies suggested that surface electrostatics affect the binding and cleavage process as well as it also has an impact over stability and off-target activity. Since, the bound sgRNA and tDNA reside majorly in the REC lobe, thus studying the charge distribution of the the the REC lobe among different orthologs could provide some insights behind the specificity of each ortholog. The structure of REC lobes of eight Cas9s were extracted for every ortholog of Cas9 and their surface electrostatic potential map was generated using command line tool APBS version 1.5 Baker *et al.* (2001). The output generated for every ortholog was visualised using PyMol Schrodinger and DeLano (2020*a*). The Figure1.9B shows a surface electrostatic potential map of REC lobes of SpCas9 orthologs colored from -5 kT/e (red) to +5 kT/e (blue). This figure supports our finding from domain architecture Figure1.9A, that the

length/size of REC lobe varies a lot as well as it also tells that the cavity of sgRNA/tDNA binding is predominantly positively charged, neutralizing the negative charge of nucleic acids. The REC lobe of FnCas9 (PDB code-5B2Q) appears to have highly positive electrostatic potential spread at a larger surface area. A work by Nishimasu et al., suggested that the REC lobe is one of the least conserved regions across the three Cas9 families within the type II CRISPR system (IIA, IIB, and IIC), and many Cas9 orthologs contain significantly shorter REC lobes Nishimasu *et al.* (2014*a*). In this structural analyses, FnCas9 stands out due to its longer REC lobe and high electrostatic potential compared to other orthlogs enhancing its ability to accommodate nucleic acids while maintaining specificity. Furthermore, structural studies indicate that the REC lobe is one of the least conserved regions across Cas9 families, suggesting that the unique features of FnCas9 may provide advantages in terms of specificity and efficiency Nishimasu *et al.* (2014*b*).

FnCas9 possesses a novel enzymatic property that cleaves the target DNA in a staggered pattern to leave 4-nt 5′-overhangs and exhibits a higher intrinsic specificity as compared with SpCas9 Acharya *et al.* (2019*b*). Specifically, FnCas9 can only tolerate a mismatch on the 5′-end position distal to the PAM, whereas SpCas9 can tolerate the mismatch on six positions, one of which was proximal to the PAM, and a high level (31%) of indels was induced on the mismatched target. Furthermore, on the 5′-end mismatch position, the FnCas9 nuclease activity was only 35% of the SpCas9 activity. As per the speculations, the possible reason behind the higher efficacy of FnCas9 could be the higher electrostatic potential of FnCas9 than SpCas9, interacting with several bases, or the presence of a structural zinc ion in a coordination sphere consisting of 4 cysteine residues, leading to less tolerance to mismatches by FnCas9 Chen *et al.* (2017*a*). The unique structural features of FnCas9, including its electrostatic properties and domain organization, might contribute to its stringent binding characteristics, making it less tolerant to mismatches in the guide RNA. While the precise factors responsible for FnCas9's higher specificity remain largely unknown, molecular dynamics (MD) simulations can provide valuable insights into its binding interactions and cleavage mechanisms. FnCas9 comprises seven domains: REC1, REC2, REC3, RuvC, HNH, WED, and PI. Notably, the REC2 domain is inserted into the REC1 domain, and the REC1 and REC3 domains are linked by a REC1–REC3 linker (Figure 1.10A). The RuvC domain interacts with the REC3 domain rather than the PI domain, which instead interacts with the WED domain. This arrangement contrasts with SpCas9, which features a bilobed architecture consisting of REC and NUC lobes. In SpCas9, the RuvC domain interacts with the PI domain to form a platform for binding the 3′ tracrRNA tail. The structure of FnCas9's REC2 adopts a new fold that is structurally unrelated to that of SpCas9. These differences in inter-domain interactions and structural arrangements contribute to their distinct functionalities and targeting mechanisms in genome editing applications (Figure 1.10).

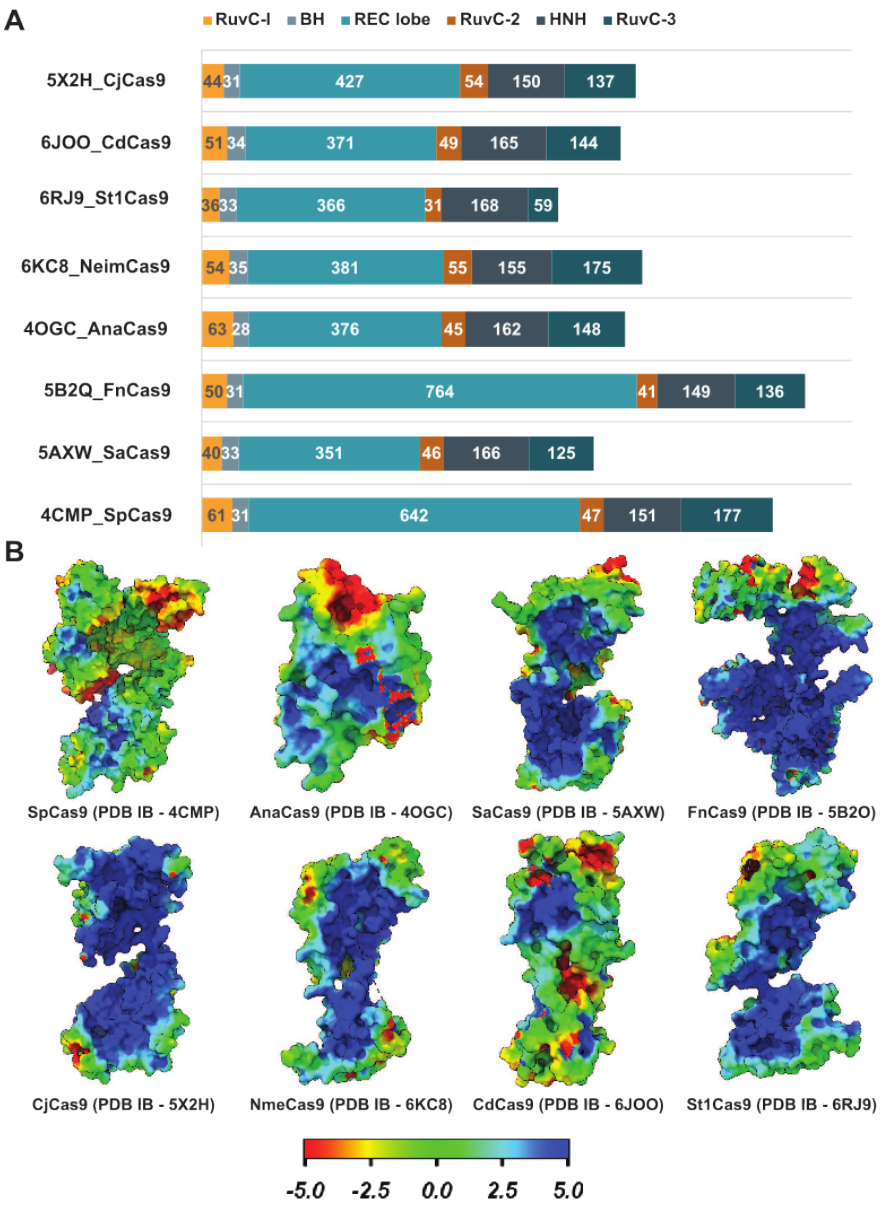


Figure 1.9: A. Domain architecture of Cas9 orthologs: Organisation of NUC lobe (HNH and RuvC domains) and REC lobe among different orthologs of Cas9 is shown. RuvC-I, II, and III (deep yellow, mustard yellow, and dark green), HNH domain (dark-grey), BH-Bridge Helix (grey), and REC lobe (green color). The PDB ID and species name of each ortholog (right side) and length of each domain are written in each box. B. Surface electrostatic potential map of REC lobes of Cas9 orthologs colored from −5kT/e (red) to +5kT/e (blue).

1.7 Thesis Objectives and Organization

SpCas9, derived from the bacterium Streptococcus pyogenes is the most commonly used DNA cleaving enzyme from the most well-characterized, type-II CRISPR/Cas system. Despite the growing utilization of the CRISPR/Cas9 system in genome engineering, some limitations hinder its widespread application. This includes the large size of Cas9 (difficult in delivery), recognition of a specific PAM (limiting its effectiveness), and introduction of unintended genetic modifications of sequences similar to those of target genes. The site-specific cleavage by CRISPR/Cas

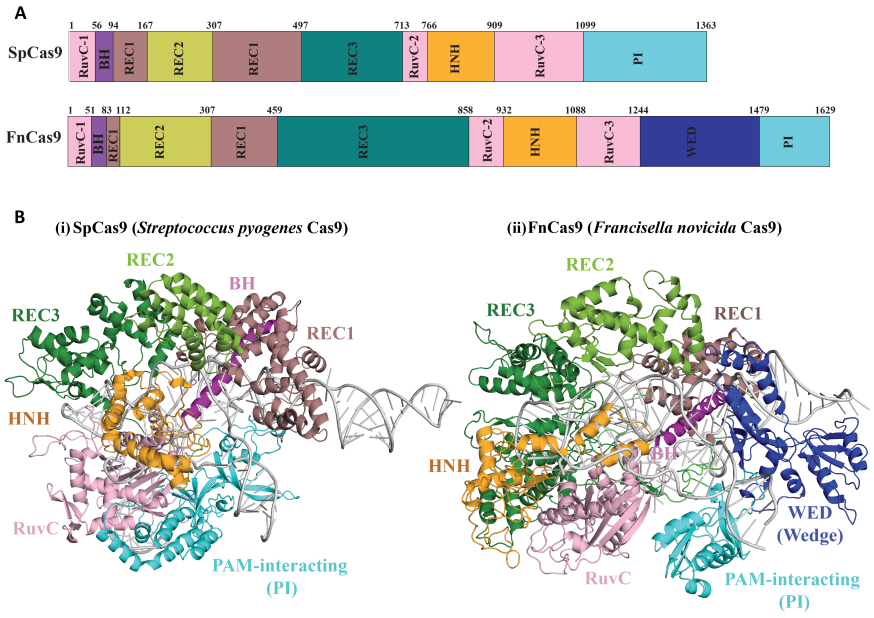


Figure 1.10: A. Structural and sequence comparison of SpCas9 and FnCas9. (A) Domain architecture of SpCas9 (Streptococcus pyogenes Cas9) and FnCas9 (Francisella novicida Cas9), showing the key functional domains and their residue boundaries.

system fails 15% of the times Zhang *et al.* (2015*a*); Cho *et al.* (2014*a*); Wu *et al.* (2014), leading to emergence of novel high-fidelity Cas variants (i.e., SuperFi-Cas9, Hifi Cas9, eSpCas9(1.1), SpCas9-HF1, SpCas9-NG, xCas9) with improved efficiency and ease of delivery Manghwar *et al.* (2019); Casini *et al.* (2018*a*); Lee *et al.* (2018*a*); Vakulskas *et al.* (2018*a*); Bravo *et al.* (2022*b*). Although, Cas9 variants with improved discriminating against mismatches have been developed, these suffer from significantly lower rates of on-target DNA cleavage. FnCas9, a naturally occurring Cas9 ortholog has been reported to have less tolerance to mismatches. Extensive structural and mechanistic studies involving atomistic simulations, targeted-MD, and accelerated-MDperformed for SpCas9 provided interesting insights into understanding the mode of action and kinetics behind CRISPR/Cas9 mediated DNA cleavage. The power of domain-rearrangement in cleavage activity to nucleic-acid binding has been covered in several cryo-EM, smFRET, and molecular dynamic simulation-based studies, a crosstalk between these findings has been discussed, which could help in deciphering the structural changes associated with DNA-cleavage mechanism Jiang *et al.* (2016); Huai *et al.* (2017); Osuka *et al.* (2018). Atomistic and accelerated simulations have been performed by several groups to understand the off-target effects of Cas9 and deduce it's mechanism by using QM/MM simulations Ricci *et al.* (2019). However, the exact atomistic understanding and underlying kinetics behind DNA cleavage process by more tolerant Cas9, FnCas9 remains elusive. Questions pertaining to the molecular inter-play for the interactions and a comparison between the two orthologs FnCas9 with SpCas9 can provide us insights into the elusive future of designing customized Cas9 molecules. A comparative structural and dynamic study of FnCas9 with SpCas9 can help in understanding the FnCas9 protein's structural dynamics, cleavage mechanisms, protein-DNA interactions, and optimizing its functionality.

This thesis focuses on deciphering structural factors aiding higher specificity of Cas9 by ap-

plying molecular dynamic simulations and machine-learning techniques. Deducing the structural and mechanistic insights of the FnCas9 system will not only provide us its underlying mechanism of DNA cleavage but will also be crucial for comprehending the role of electrostatics and optimal DNA binding time in regulating off-target effect in revealing the physico-chemical reason behind increased specificity and sensitivity of FnCas9 than SpCas9, thereby expanding the applicability and design of better and more specific CRISPR-Cas9.

The thesis is organized into seven chapters:

Chapter 1 provides an introduction and rationale for the study, outlining its significance in understanding FnCas9's high specificity and potential genome-editing applications. **Chapter 2** presents the background and methodologies used across the thesis objectives, describing techniques central to the structural, dynamic, and machine learning analyses of Cas9 systems. **Chapters 3 to 6** cover the research objectives in detail, each focusing on specific aspects of FnCas9's structural and mechanistic properties, comparative studies with SpCas9, engineered variants, and interactions with different nucleic acid substrates. **Chapter 7** concludes the thesis, summarizing the key findings and proposing directions for future research to further advance CRISPR-based gene-editing technologies.

The objectives of this thesis are as follows (Figure 1.11):

Objective 1: Comparative Structural and Dynamics Study of Free and gRNA-bound FnCas9 and SpCas9 Proteins This objective aims to explore the differences in structure and dynamics between FnCas9 and SpCas9, focusing on how gRNA binding affects each protein. FnCas9 and SpCas9, although homologous, exhibit different behaviors and cleavage mechanisms when targeting DNA sequences. To understand these distinctions, molecular simulations, such as Molecular Dynamics (MD) simulations, are employed. By simulating both the apo (unbound) and gRNA-bound states of each protein, we observe how binding impacts the conformational landscape and domain organization of these proteins. Special focus is given to the electrostatic interactions and the role of critical structural elements, like the bridge-helix, which may help stabilize FnCas9 in a cleavage-competent state. Through Principal Component Analysis (PCA) and Free Energy Landscape (FEL) mapping, we assess shifts in protein dynamics and flexibility, shedding light on the structural adaptations that may underlie FnCas9's higher specificity. This analysis provides insights into domain rearrangements that make FnCas9 less prone to off-target effects, which is crucial for enhancing its use in genome editing.

Objective 2: Deciphering Cas9 Specificity: Role of Domain Dynamics and RNA:DNA Hybrid Interactions Revealed through Machine Learning and Accelerated Molecular Simulations The second objective centers on FnCas9's specificity and its ability to discriminate between on-target and off-target heteroduplexes. Specifically, it examines the protein's response to base-pair mismatches in the RNA:DNA hybrid. To mimic potential off-target conditions, mismatches are introduced at specific positions distal to the PAM (Protospacer Adjacent Motif) in the ternary complexes of FnCas9 and SpCas9. The response to these mismatches is analyzed using a combination of accelerated molecular dynamics (GaMD) and machine learning models, which reveal

unique structural factors associated with base-pair discrimination. By comparing conformational changes and identifying areas of structural rigidity or flexibility, this objective aims to pinpoint the specific features responsible for the high base-discrimination capability of FnCas9. Additionally, Markov State Modeling (MSM) captures distinct conformational states that reflect the pathway of the mismatch recognition process. The metastable identified from MSM were used as labels and protein and RNA:DNA hybrid structural parameters were used features to predict the metstable states of the RNA:DNA hybrid. Feature ranking methods were applied to determine key features which aided in the classification of the heteroduplex states. MM/GBSA method was used to find the binding affinity of REC and NUC lobes with the hybrid in SpCas9 and FnCas9. These methods and the observations made furthered our understanding the structural factors responsible for specificity of FnCas9 than SpCas9.

Objective 3: Structural Insights into Mismatch Recognition by Engineered FnCas9 Variants This objective addresses the optimization of FnCas9 for improved targeting efficiency and specificity, essential for therapeutic applications. Unlike wild-type FnCas9, engineered variants are modified to enhance PAM compatibility and cellular delivery. Previous studies have shown that wild-type FnCas9's cellular targeting can be limited; therefore, variants with expanded PAM specificity and truncated domain structures have been developed to overcome these barriers. By comparing wild-type FnCas9 to engineered versions, including those with a truncated REC2 domain in their on-target and off-target forms, this objective seeks to identify how domain removal or modification impacts protein dynamics and binding affinity. Molecular dynamics simulations of these variants reveal how the absence of the REC2 domain or modifications in key interaction networks can lead to similar cleavage efficiency while potentially increasing the protein's applicability for precise gene editing. Additionally, the structural changes in the REC2-lacking variants provide insights into the protein's adaptability, helping to optimize FnCas9 for various cellular contexts and broaden its therapeutic potential.

Objective 4: Expanding FnCas9 Functionality: Comparative Analysis of RNA and DNA Targeting The final objective focuses on FnCas9's potential to interact with substrates other than DNA, specifically tRNA. Understanding these interactions could open new avenues for RNA-based applications of CRISPR technology. By substituting target DNA (tDNA) with tRNA in the FnCas9 complex, this objective aims to map out how FnCas9 interacts with RNA and identify the key residues involved in binding. Atomistic simulations provide insight into whether tRNA binds through specific residues or shares common binding sites with tDNA. The analysis extends to comparing the binding affinities of tRNA and tDNA, examining the influence of WED-PI domain interactions on substrate recognition. Through simulations, the conformational changes induced by tRNA binding are captured, allowing us to understand the mechanistic differences between RNA and DNA interactions with FnCas9. These insights could lead to engineered Cas9 variants with enhanced affinity and efficiency for RNA targeting, broadening the scope of CRISPR applications in RNA editing and regulation.

The goal of this thesis is to advance the understanding and application of CRISPR/Cas9 technology by exploring the structural and mechanistic factors that contribute to the higher specificity

of FnCas9, particularly in comparison to SpCas9. Through a detailed study of FnCas9's unique structural features and domain dynamics when bound to guide RNA (Objective 1), the work investigates how these factors influence DNA binding and cleavage. By analyzing how FnCas9 and SpCas9 respond to base-pair mismatches in the RNA hybrid (Objective 2), the research provides insight into FnCas9's enhanced ability to discriminate between target and off-target sequences. The thesis also examines engineered FnCas9 variants designed for improved efficiency and cellular delivery, identifying structural adaptations that make them more suitable for therapeutic applications (Objective 3). Lastly, it explores FnCas9's interactions with tRNA to assess the potential for RNA targeting, offering insights into new CRISPR applications (Objective 4). Together, these objectives aim to expand the versatility and precision of CRISPR-Cas9, laying a foundation for more specific, efficient genome-editing tools.

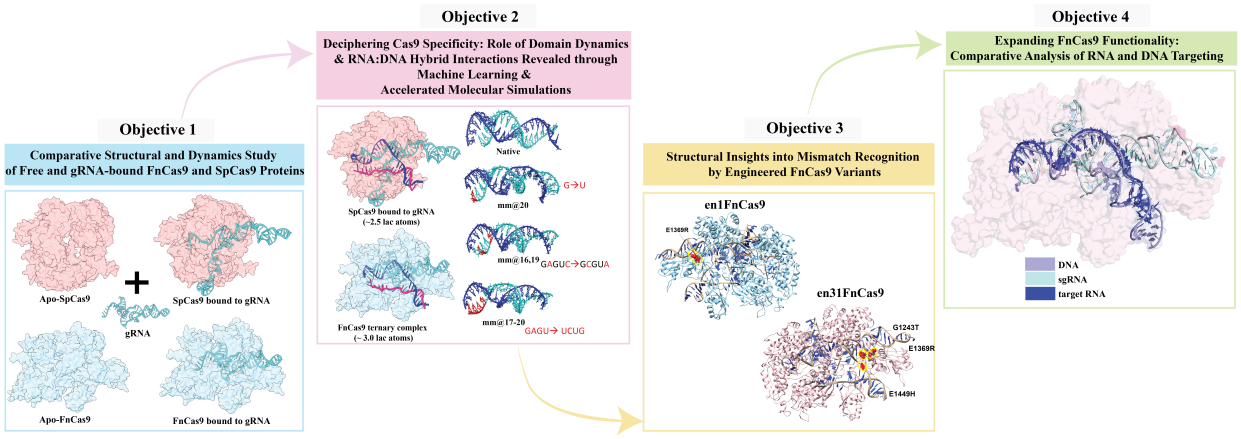


Figure 1.11: Thesis Organisation: The thesis comprises seven distinct chapters, including the introduction chapter, a chapter on the background and related literature, four chapters on each objective designed for deciphering the key structural determinants of CRISPR/Cas9, last chapter talks about the conclusion and future work.

CHAPTER 2

Background

2.1 Molecular Dynamics and its Applications on the CRISPR/Cas9 System

2.1.1 Molecular Dynamics

Molecular dynamics (MD) simulation is a powerful computational method used to study the atomic and molecular interactions that govern the behavior of physical systems at both microscopic and macroscopic scales. Based on the principles of statistical mechanics and classical dynamics, MD simulations allow researchers to explore the dynamic evolution of molecular systems over time. The potential energy of the system is derived from both bonded and non-bonded interactions, with the parameters used in these calculations referred to as a "force field." Commonly used force fields include CHARMM Vanommeslaeghe *et al.* (2010), GROMOS Oostenbrink *et al.* (2004*a*), and AMBER Weiner *et al.* (1986), each with its own parameterization methods that affect the accuracy and applicability of the simulations. One of the primary advantages of MD simulations is their ability to probe vast spatiotemporal landscapes, enabling the examination of systems with thousands of atoms over timescales that can reach milliseconds. This capability is particularly valuable in the study of complex biological macromolecules, such as proteins and nucleic acids, where understanding their dynamic properties is crucial for elucidating structure-function relationships. The evolution of MD simulations has been facilitated by advances in high-performance computing, allowing for the routine simulation of systems containing 50,000 to 500,000 atoms Hospital and et al. (2015).

Classical Mechanics : Equations of Motion

MD simulations operate by numerically solving Newton's equations of motion for interacting particles, where the forces between atoms are calculated using inter-atomic potentials defined by the chosen force field.

$$\mathbf{F}_i = m_i \mathbf{a}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -\frac{dV}{dr_i} \tag{2.1}$$

where $\mathbf{F}_i$ is the force exerted on particle i, m_i is the mass of particle i,V is the potential energy of the system and $\mathbf{a}_i$ is the acceleration of particle i. Therefore, to calculate a trajectory, one only needs the initial positions of the atoms, an initial distribution of velocities and the acceleration,

which is determined by the gradient of the potential energy function. The equations of motion are deterministic, e.g., the positions and the velocities at time zero determine the positions and velocities at all other times, t. The initial coordinates for atoms can be obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) website Berman *et al.* (2000) which provides the Protein Data Bank (PDB) files for experimentally resolved structures of biomolecules. These structures have been resolved using various experimental techniques, such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. The initial distribution of velocities are usually determined from a random distribution with the magnitudes conforming to the required temperature and corrected so there is no overall momentum, i.e.,

$$P = \sum_{i=1}^{N} m_i v_i = 0 \tag{2.2}$$

The velocities, vi, are often chosen randomly from a Maxwell-Boltzmann or Gaussian distribution at a given temperature, which gives the probability that an atom i has a velocity v_x in the x direction at a temperature T.

$$P(v) = \sqrt{(\frac{m}{2\Pi\kappa_B T})^3 4\Pi} \mathrm{e}^{-\frac{mv^2}{2\Pi\kappa_B T}} \tag{2.3}$$

where *m* is the mass of the a particle, *v* is the velocity, κ_B is the Boltzmann constant, and T is the temperature.

2.1.2 Ensembles

In molecular dynamics (MD) simulations, the thermodynamic state of a system is characterized by macroscopic properties such as temperature (T), pressure (P), and volume (V). These macroscopic properties are linked to the microscopic state, which details the coordinates and velocities of each particle in the system. To bridge this gap, the concept of an ensemble is utilized, representing a collection of weighted microstates that correspond to a single macrostate. Essentially, one macrostate can be associated with numerous microstates. Different ensembles are defined based on which thermodynamic properties are held constant.

- Microcanonical ensemble (NVE) maintains a constant number of particles (N), volume (V), and energy (E). However, most chemical and biological processes occur under constant temperature and pressure conditions, making the canonical (NVT) and isothermal-isobaric (NPT) ensembles more relevant for simulations Leach (2001).
- In the NVT ensemble, the number of particles, volume, and temperature are fixed, allowing for heat exchange with the surroundings through a thermostat Berendsen *et al.* (1984).
- Conversely, the NPT ensemble maintains constant particle number, pressure, and temperature, requiring both a thermostat and a barostat for pressure control Martyna *et al.* (1994). Most chemical and biological processes occur at constant temperature and constant pressure. Therefore, by using a thermostat or/and a barostat algorithm, the canonical ensemble (NVT) or the isothermal-isobaric ensemble (NPT) can be sampled in MD simulations Barth and Schlick (1998).

Thermostats

In order to obtain a realistic reproduction of the biological conditions, constant temperature can be obtained by coupling the system to a thermostat. The main used methods are: (i) the weak coupling method (Berendsen thermostat) Berendsen *et al.* (1984) and (ii) the coupling to an external bath (Nosé Hoover thermostat) Nose and Klein (1986) In the Berendsen thermostat, the temperature of the system is kept close to the target equation by the equation where T is the instantaneous value of the temperature and T_τ is the temperature coupling time.

$$\frac{dT}{dt} = \frac{1}{T_t}[T_0 - T_\tau] \tag{2.4}$$

The kinetic energy of the system is given by:

$$E_{\text{kin}}(t) = \sum_{k=1}^{N} m_k \dot{q}_k^2(t) = \frac{1}{2} N_{\text{df}} k_B T(t) \tag{2.5}$$

The atomic velocities can be scaled by a factor λ:

$$\delta E_{\text{kin}}(t) = \left[\lambda(t)^2 - 1\right] \frac{1}{2} N_{\text{df}} k_B T(t) \tag{2.6}$$

The factor $\lambda(t)$ is used to scale the velocities $\dot{q}_k$ at each step of integration, so as to relax the temperature towards the target value T_0. The time coupling constant T_τ controls the relaxation rate. The Nosé-Hoover thermostat Nose and Klein (1986) extends the ensemble via the introduction of a thermal reservoir and a friction term in the equations of motion. The friction force is proportional to the product of the velocity of each particle and a friction parameter ξ. The ξ parameter is a dynamic quantity with its equation of motion. The time derivative is calculated from the difference between the current kinetic energy and the reference temperature. The equation of motion of each particle becomes:

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i} - \xi \frac{dr_i}{dt} \tag{2.7}$$

whereas the equation of motion for the bath parameter ξ becomes:

$$\frac{d\xi}{dt} = \frac{1}{Q}(T - T_0) \tag{2.8}$$

where T_0 is the reference temperature and Q controls the strength of coupling to the heat bath.

Barostats

The pressure of the system can be controlled by coupling the system to a barostat, such as the method of Berendsen Berendsen (1999). This method scales the coordinates and the box vectors, obtaining a first–order kinetic relaxation of the instantaneous pressure *p(t)* toward the reference

pressure *p0* with a time constant τ_p. Similarly to the Berendsen thermostat:

$$\frac{dp(t)}{dt} = \frac{p_0 - p_t}{\tau_p} \tag{2.9}$$

the scaling factor μ(t) for atomic positions is:

$$\mu_{\alpha\beta} = \delta_{\alpha\beta} - \beta_{\alpha\beta}\frac{\Delta t}{3\tau_p}[p_{0,\alpha\beta} - p_{\alpha\beta}(t)] \tag{2.10}$$

where β is the isothermal compressibility, p_t is the instantaneous pressure, p_0 is the target pressure and τ_p is the pressure coupling time constant. Key features include that the Berendsen Barostat continuously rescales the simulation box to match the target pressure. A parameter called the *pressure coupling constant* controls the relaxation time over which pressure changes take place. It applies a damping effect to the pressure, which makes it good for quickly bringing the system to the desired pressure but less accurate in producing correct fluctuations of pressure. However, it does not generate a true NPT ensemble, meaning the fluctuations in pressure are artificially suppressed. While it is effective for equilibrating a system, it does not give correct pressure distribution statistics.

Another method, the Parrinello-Rahman pressure coupling , allows for fluctuations in volume and shape, using an extended Lagrangian that incorporates additional coordinates and velocities. This method is often combined with the Nose-Hoover thermostat Nose and Klein (1986) for enhanced control. Unlike the Berendsen barostat, which only changes the volume isotropically (i.e., the box expands or contracts uniformly in all directions), the Parrinello-Rahman barostat allows the box to change its shape, which is particularly useful when studying phase transitions or systems under anisotropic stress. It generates a true NPT ensemble, meaning that the fluctuations in pressure and volume correspond to the correct thermodynamic ensemble, giving better accuracy for pressure-related quantities. It is computationally more expensive than the Berendsen barostat because of its complexity in handling the shape of the box and the need to solve additional equations of motion. The Parrinello-Rahman barostat modifies the MD equations by coupling the system to a *tensor of stresses*. The equations of motion for the box vectors h are as follows:

$$M_h\frac{d^2h}{dt^2} = P - P_0 \tag{2.11}$$

Where M_h is the "mass" associated with the box matrix, h is the matrix of box vectors defining the simulation cell, P is the instantaneous internal pressure tensor P_0 is the external pressure tensor. Both methods aim to replicate realistic conditions in simulations by adjusting the system's volume and pressure dynamically, ensuring that the molecular dynamics reflect the desired thermodynamic ensemble. By implementing these techniques, researchers can achieve more accurate simulations that closely mimic experimental conditions, facilitating a deeper understanding of molecular behavior in various environments.

2.1.3 Integration Algorithms

After defining the initial coordinates (position) and velocities of the particles, we need to perform numerical integration of the Newton's equations of motion. The potential energy is a function of the atomic positions (3N) of all the atoms in the system. Due to the complicated nature of this function, there is no analytical solution to the equations of motion; they must be solved numerically. Numerous numerical algorithms have been developed for integrating the equations of motion Ryckaert *et al.* (1977). All the integration algorithms assume the positions, velocities and accelerations can be approximated by a Taylor series expansion:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + ... \tag{2.12}$$

$$v(t+\delta t) = v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + ... \tag{2.13}$$

$$a(t+\delta t) = a(t) + b(t) + ... \tag{2.14}$$

Where r is the position, v is the velocity (the first derivative with respect to time), a is the acceleration (the second derivative with respect to time), etc. Some of the commonly used integration algorithms are -

- *Verlet algorithm*: The Verlet algorithm uses positions and accelerations at time t and the positions from time t-δt to calculate new positions at time t+δt. The Verlet algorithm uses no explicit velocities. The advantages of the Verlet algorithm are, i) it is straightforward, and ii) the storage requirements are modest. The disadvantage is that the algorithm is of moderate precision Verlet (1967).
- *Leap-frog algorithm* : In this algorithm, the velocities are first calculated at time t+1/2dt; these are used to calculate the positions, r, at time t+dt. In this way, the velocities leap over the positions, then the positions leap over the velocities. The advantage of this algorithm is that the velocities are explicitly calculated, however, the disadvantage is that they are not calculated at the same time as the positions. The velocities at time t can be approximated by the relationship Hockney (1970).
- *Velocity Verlet*: This algorithm yields positions, velocities and accelerations at time t. There is no compromise on precision Swope *et al.* (1982).
- *Beemans algorithm*: This algorithm is closely related to the Verlet algorithm. The advantage of this algorithm is that it provides a more accurate expression for the velocities and better energy conservation. The disadvantage is that the more complex expressions make the calculation more expensive Beeman (1976).

The selection of an appropriate time step (Δt) for numerical integration is crucial in molecular dynamics (MD) simulations, as it directly impacts the accuracy and convergence of the results. A smaller time step enhances the accuracy of the simulations but increases computational costs, while a larger time step enables more extensive sampling of the conformational space but may lead to instabilities in the simulation. Therefore, it is essential to choose the optimal time step to achieve a balance between accuracy and computational efficiency. One key requirement for

numerical integrators is that the time step should be sufficiently small compared to the most rapid component of the motion in the system Phillips *et al.* (2005). For MD simulations of biomolecules, where the fastest motions typically involve hydrogen atoms, a recommended time step is either 1 fs if the bonds to hydrogen atoms are flexible or 2 fs if the bonds to hydrogen atoms are constrained using rigid bond algorithms Schlick *et al.* (1997).

2.2 Molecular Dynamic Simulation of Proteins-Nucleic Acid Complexes

2.2.1 Force-Field

In classical MD the potential energy is determined by an empirical force field (FF) parameterized in order to reproduce experimental or ab-initio data. In order to determine the electronic energy for a given nuclear configuration of an atomic system, quantum mechanical calculations are required. In FF-based methods, the potential energy is calculated as a parametric function of the nuclear. Several force fields have been developed for biomolecular applications such as AMBER Cornell *et al.* (1995), GROMOSOostenbrink *et al.* (2004*a*) and CHARMM Vanommeslaeghe *et al.* (2010). In this thesis, we use the Amberff14SB Maier *et al.* (2015) and its recent ff99bsc0 corrections for DNA Pérez *et al.* (2007) and the ff99bsc0+ χOL3 corrections for RNA Banás *et al.* (2010); Zgarbov *et al.* (2011). Its pairwise-additive potential is of the form:

$$\begin{aligned} V = & \sum_{\text{bonds}} K_R(R - R_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{\text{eq}})^2 \\ & + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ & + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \end{aligned} \tag{2.15}$$

The Amber ff14SB force field is widely recognized for its accuracy and reliability in biomolecular simulations, offering significant improvements in the representation of protein side-chain conformations and backbone dynamics compared to earlier AMBER force fields Maier *et al.* (2015). These enhancements are particularly important for capturing the subtle conformational transitions relevant to protein-nucleic acid interactions. For nucleic acids, the ff99bsc0 correction (Pérez et al., 2007) for DNA and the χOL3 correction for RNA Banás *et al.* (2010); Zgarbov *et al.* (2011) have been shown to substantially improve the description of backbone torsion angles and overall helical stability. The χOL3 modification, in particular, addresses the glycosidic torsion angle in RNA, yielding more accurate structural ensembles and better agreement with experimental data. These corrections are crucial for systems involving RNA:DNA hybrids and structured RNAs, as in our study. While alternative force fields such as CHARMM Best *et al.* (2012) and GROMOS Oostenbrink *et al.* (2004*b*) are also established choices in biomolecular simulations, the Amber

ff14SB suite, combined with the aforementioned nucleic acid corrections, has demonstrated excellent performance in modeling protein-nucleic acid complexes. This is supported by numerous benchmarking studies showing its ability to reproduce experimental structures and dynamics for both proteins and nucleic acidsLindorff-Larsen *et al.* (2012); Bussi *et al.* (2007); Piana *et al.* (2013); Zgarbová and et al. (2015); Bergonzo and Cheatham III (2015)

The interactions are divided into bonded interactions, acting within a molecule, and non-bonded interactions, acting between all atoms except from bonded neighbours. Eq. 2.15 contains three terms for the bonded interactions, namely one for the chemical bonds between two neighboring atoms, one for the bond angles between three atoms, and one for the dihedral angles between four atoms. Additionally, improper dihedral-angle terms can be applied to maintain planar or tetrahedral conformations. The functional form of the bond and angle terms is quadratic, while the dihedral term uses a trigonometric function. In Eq. 2.15, R is the distance between two atoms i and j that are bound together via a covalent bond, R_{eq} and θ_{eq} refer to equilibrium bond lengths and angles, K_R and K_θ are the vibrational constants. V_n is the torsional barrier corresponding to the n^{th} barrier of a given torsional angle with phase γ. The last term in Eq. 2.15 refers to the non-bonded interactions, which are composed of a Lennard-Jones term for the van-der-Waals interactions and a Coulomb term for the electrostatic interactions between atoms i and j. Van der Waals and Coulomb interactions between atoms i and j that are involved in direct (1-2) or indirect (1-3, 1-4) bonded interactions are excluded from the potential or rescaled.

Eq.2.16 reports the van der Waals term that describes the repulsive forces at small interatomic distances due to the Pauli repulsion between electrons (decaying exponentially, modeled with R_{ij}^{-12}), and the attractive forces at intermediate distances due to instantaneous dipole-induced dipole interactions (decaying with R^{-6}).

$$V_{\text{vdW}} = \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right] \tag{2.16}$$

The Coulomb term eq 2.17 takes into account the electrostatic interactions between charged particles.

$$V_{\text{el}} = \sum_{i<j} \frac{q_i q_j}{\epsilon R_{ij}} \tag{2.17}$$

The calculation of the electrostatic Coulomb interactions between charged particles is computationally the most challenging as it represents a long-range interaction and converges very slowly. In the AMBER/parm99 FF, the electrostatic properties of the electron density of the system are reproduced by atom-centred point charges that are derived from quantum chemical calculations within a restrained electrostatic potential (RESP) fitting scheme Wang *et al.* (2004); Hoover (1985). Non-bonded interactions represent a major part of the computational cost, and their treatment requires special care. To begin with, in order to avoid artifacts due to finite-size effects, the boundary is usually completely removed by treating the system with periodic boundary conditions (PBC). This

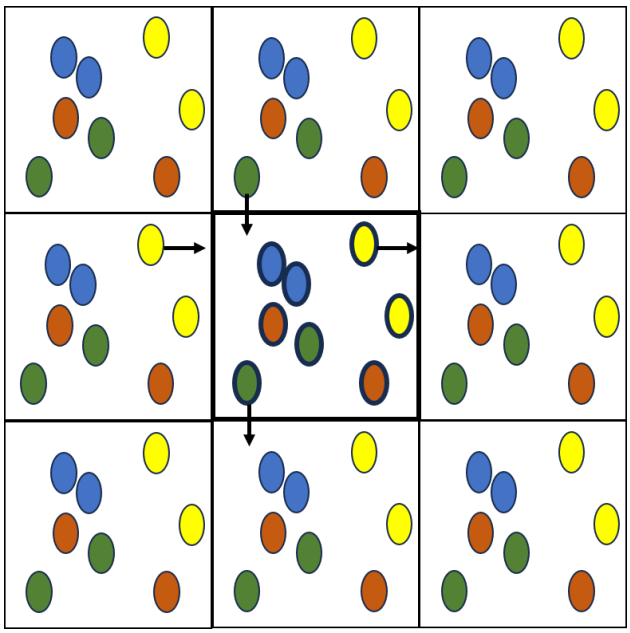


Figure 2.1: In Periodic boundary consitions, the simulation box is surrounded by identical copies of itself in all directions (x, y, and z axes). Whenever a molecule in the original box moves out of one side, an equivalent molecule enters the box from the opposite side.

places the system in an environment similar to an infinite solution. The most widely used algorithm for a fast and accurate treatment of the Coulomb term makes use of the exact periodicity of the system and includes all electrostatic interactions (also with periodic copies) by lattice sum techniques based on Ewald summation, among which the particle mesh Ewald (PME) method is widely used Ewald (1921). Despite the high reliability of modern FFs, they suffer from intrinsic limitations like limited transferability, neglect of polarisation and their inability to describe chemical bond breaking and formation.

2.2.2 Boundary Condition

The treatment of boundaries is a crucial aspect of molecular dynamics (MD) simulations, particularly because typical simulated systems are significantly smaller than macroscopic matter, often comprising around 10^4 to 10^5 number of atoms. When molecules are arranged in a cubic box, a substantial number of them will be located at the surface, experiencing different forces compared to those in the bulk. This finite size can lead to boundary effects that may significantly influence the results, especially when simulating homogeneous liquids or solutions Hess (2008). To mitigate these surface effects, periodic boundary conditions (PBC) are commonly employed (shown in Figure 2.1). This method involves simulating the system within a central cubic box surrounded by an infinite number of identical copies. As a result, when a molecule exits the central box, one of its periodic images enters from the opposite side, effectively eliminating physical boundaries and surface molecules Allen and Tildesley (2017).

2.2.3 Enhanced Efficiency Methods

Since the first application of MD to biomolecular systems over 20 years ago, various methods have been developed to enhance the efficiency and timescales of MD simulations. The integration of Newton's equations of motion is limited by the highest frequency vibrations in the system, particularly those involving hydrogen atoms, which typically have bond stretching frequencies around 10^4 Hz. This limitation restricts the time step in MD simulations to approximately 0.5 fs. However, methods such as the SHAKE algorithm, which constrains bond lengths, allow for an increased time step of about 2 fs without significantly affecting the overall dynamics of the system Hess (2008). While bond-angle constraints can impact other fluctuations, the identification of distinct classes of fluctuation frequencies enables the design of more efficient algorithms Kräutler *et al.* (2001). Rapidly fluctuating forces can be recalculated more frequently than those with longer timescales. Multiple time-step algorithms have been successfully implemented, achieving speed-up factors of 4-5 compared to unconstrained dynamics, making them slightly more efficient than simulations with bond-length constraints. Another approach involves constraining rotational and translational motions in polyatomic systems, which is particularly useful for studying biomolecules in solution. This method, often coupled with SHAKE in a leap-frog integration scheme, reduces molecular relaxation time and allows for a decrease in the number of solvent molecules needed Bekker and et al. (2004)

2.2.4 Enhanced Sampling Methods

Enhanced sampling methods in molecular dynamics (MD) simulations are essential for efficiently exploring the vast conformational space of complex molecular systems. Traditional MD simulations often struggle to sample rare events or transitions due to high energy barriers separating different states. Enhanced sampling techniques address this limitation by modifying the potential energy landscape or introducing external biases, allowing for a more effective exploration of relevant regions in configuration space.Several enhanced sampling methods have been developed, including:

1. **Metadynamics**: This technique involves adding a history-dependent bias to the potential energy surface, which helps to overcome energy barriers and encourages the sampling of previously unvisited states. It is particularly effective in studying free energy landscapes and reaction pathways Bernardi *et al.* (2014).

2. **Umbrella Sampling**: This method utilizes a series of overlapping simulations at different values of a collective variable (CV) to obtain a comprehensive view of the free energy landscape. By applying biasing potentials, umbrella sampling allows for the efficient sampling of states that might be rarely visited in standard MD Torrie and Valleau (1977).

3. **Replica Exchange Molecular Dynamics (REMD)**: This technique involves running multiple simulations at different temperatures or conditions and periodically exchanging configurations between them. This approach enhances sampling by allowing the system to escape local minima more easily Sugita and Okamoto (2000).

4. **Adaptive Biasing Force (ABF)**: ABF methods apply a biasing force that adapts based on the system's response, allowing for efficient sampling of the free energy landscape without the need for prior knowledge of the system's behavior Darve *et al.* (2008).

5. **Gaussian Accelerated Molecular Dynamics (GaMD)**: GaMD enhances the sampling of rare events by applying a Gaussian bias to the potential energy surface, which reduces the energy barriers that the system must overcome. This method is beneficial for studying processes that occur on timescales that are typically inaccessible to conventional MD simulations, such as protein folding or ligand binding Klein and et al. (2018). GaMD is especially advantageous because it can be easily integrated into existing MD frameworks and does not require extensive parameter tuning, making it user-friendly and versatile for a wide range of applications.

Gaussian Accelerated MD simulations

Gaussian Accelerated MD simulations are state-of-the-art "accelerated" MD which uses a quadratic functions to decrease the potential energy barriers and accelerate the transitions between low-energy states. It samples a wider configuration space and capture biological processes that cannot be described via conventional MD simulations. Given the complexity of the CRISPR-Cas9 system, Accelerated MD (AMBER22) methods represent the best approach to perform long-scale simulations. The system potential energy surface (PES) is smoothed using the improved sampling technique known as "accelerated MD" (aMD), which reduces energy barriers and speeds up transitions between low-energy states by introducing a non-negative boost potential Hamelberg *et al.* (2004). In GaMD, the boost potential applied follows a Gaussian distribution, which is a more robust and novel way of accelerating simulations Miao *et al.* (2015). This enables accurate reweighting via cumulant expansion to the second order in order to acquire the canonical ensemble by reconstructing the potential energy surface's original shape. Therefore, the same low-energy physical states are sampled in the GaMD simulations even with biasing potential. As a result, reweighting can enable the quantitative recovery of conformational distributions, whereas results that are not reweighted, such as those shown here, capture the low-energy physical states and also offer a helpful semiquantitative ranking of their probability. This technique helps to describe the canonical distribution for a large biomolecular system, such as the CRISPR/Cas9 Palermo (2019*b*); Ricci *et al.* (2019) biomolecular system, such as CRISPR/Cas9 Palermo (2019*b*); Ricci *et al.* (2019), according to available experimental data. GaMD extends the use of aMD to large biological systems, for which it has been challenging to attain the canonical ensemble, given the large statistical noise Miao and McCammon (2016). Considering a system with N atoms at positions $\vec{r} = \{\vec{r_1}, ...\vec{r_N}\}$ when the system potential $V(\vec{r})$ is lower than a threshold energy E, the energy surface is modified by adding a boost potential as

$$V^*(\vec{r}) = V(\vec{r}) + \triangle V(\vec{r}) \tag{2.18}$$

$$\triangle V(\vec{r}) = 1/2k(E - V(\vec{r}))^2 \tag{2.19}$$

where k is the harmonic force constant. The two adjustable parameters E and k are automatically determined by applying the following three criteria. First, for any two arbitrary potential values $V_1(\vec{r})$ and $V_2(\vec{r})$ found on the original energy surface, if $V_1(\vec{r}) < V_2(\vec{r}), \triangle V$ should be a monotonic function that does not change the relative order of the biased potential values, i.e., $V_1^*(\vec{r}) < V_2^*(\vec{r})$. Second, if $V_1(\vec{r}) < V_2(\vec{r})$, the potential difference observed on the smoothened energy, i.e., $V_2^*(\vec{r}) - V_1^*(\vec{r}) < V_2(\vec{r}) < V_2(\vec{r})$. By combining the first two criteria and putting in the formula of $V^*(\vec{r})$ and $\triangle V(\vec{r}$, we obtain the following result.

$$V_{max} \leq E \leq V_{min} + 1/k \tag{2.20}$$

where V_{min} and V_{max} are the system minimum and maximum potential energies. To ensure that eq 2.20 is valid, k has to satisfy $k \leq 1/V_{max} - V_{min}$. By defining $k \equiv k0(1/V_{max}) - V_{min}$, then $0 < k \leq 1$. Third, the standard deviation of $\triangle V$ needs to be small enough (i.e., narrow distribution) to ensure accurate reweighting using cumulant expansion to the second order: $\sigma_{\triangle V} = k(E - V_{avg})\sigma_V \leq \sigma_0$, where V_{avg} and σV are the average and standard deviation of the system potential energies, $\sigma_{\triangle V}$ the standard deviation of $\triangle V$, and σ_0 is a user-specified upper limit (e.g., 10 kBT) for accurate reweighting. When E is set to the lower bound, E = V_{max}, according to eq 2.20, k_0 can be calculated as

$$k_0 = min(1.0, k_0') = min\left(1.0, \frac{\sigma_0}{\sigma_V}\frac{V_{max} - V_{min}}{V_{max} - Vavg}\right) \tag{2.21}$$

Alternatively, when the threshold energy E is set to its upper bound E = $V_{min} + 1/k$, k_0 is

$$k_0 = k_0'' \equiv \left(1 - \frac{\sigma_0}{\sigma_V}\right)\frac{V_{max} - V_{min}}{V_{avg} - Vmin} \tag{2.22}$$

if k_0'' is calculated between 0 and 1. Otherwise, k_0 is calculated using eq 2.21 instead of being set to 1. The system threshold energy was set to E = V_{max} for all GaMD simulations. The boost potential has been applied in a dual-boost scheme, in which two acceleration potentials are applied simultaneously to the system: (i) the torsional terms only and (ii) across the entire potential. A time step of 2 fs was used. In my work, the maximum, minimum, average, and standard deviation values of the system potential (V_{max}, V_{min}, V_{avg}, and σ_V) were obtained from ∼50 ns cMD with no boost potential. Each GaMD simulation proceeded with a ∼25-50 ns run, in which the boost potential was updated every 1.6 ns, thus reaching equilibrium values. Finally, ∼ 500 ns of GaMD simulations were performed in the NPT ensemble. For each system, ∼500 ns (GaMD production) + 25/50 ns (GaMD equilibration) + 50 ns (pre-equilibration conventional MD) was carried out.

2.2.5 MD Simulation Trajectory Analysis

In this thesis, a comprehensive analysis of molecular dynamics (MD) simulations was conducted to investigate the structural and dynamic properties of the SpCas9 and FnCas9 proteins, particularly in their interactions with sgRNA and tDNA. The analysis included the assessment of geometrical

properties critical for understanding the system's behavior, such as the root mean square deviation (RMSD) of protein backbone atoms and RNA:DNA hybrids, solvent-accessible surface area (SASA), and the radius of gyration of Cas9. Residue-wise and Domain-wise root mean square fluctuations were computed for both Cas9 variants in their on-target and off-target forms, providing insights into the stability and flexibility of the proteins. The analysis Native contact analysis was done to quantified tight contacts within a 4 Å radius between sgRNA and tDNA within the heteroduplex, as well as native and non-native contacts formed by the heteroduplex with neighboring protein domains, including HNH, RuvC, and REC. All these analysis were performed using cpptraj Roe and Cheatham III (2013) and pyraj Roe and Cheatham (2013) module of AMBER package Case *et al.* (2023). The analysis of RNA:DNA base-pair complementarity was performed using the 3DNA code Lu and Olson (2003), focusing on various geometrical parameters that characterize base pair complementarity, such as shear, stretch, stagger, buckle, propeller, and opening. The number of native contacts formed between the sgRNA-tDNA strands was computed, along with assessments of base pairing loss and the count of non-canonical base pairs in mismatch systems. Principal component analysis (PCA) was utilized to investigate the dominant motions and conformational distributions of the RNA:DNA heteroduplex and Cas9 domains in both native forms and in the presence of base-pair mismatches. This analysis involved calculating a covariance matrix and projecting the dynamics into a low-dimensional space, allowing for a clearer understanding of the conformational changes that occur during the simulations. The linear interaction energy model was applied to calculate the electrostatic and Van der Waals interaction energies between gRNA and tDNA, providing further insights into the binding interactions and stability of the complexes.

2.3 Machine-Learning and it's application in understanding CRISPR/Cas9 Mechanism

Machine learning approaches are gaining popularity in the field of structural biology, particularly in the prediction of 3D structures and the development of machine-learned force fields. Notably, methods such as AlphaFold2 and RoseTTAFold have garnered significant attention for their ability to predict protein structures based on available structural data from the Protein Data Bank (PDB). These methods not only provide predicted structures but also generate additional information that is valuable for molecular dynamics (MD) analysis, such as defining collective variables (CVs) for enhanced sampling Team (2021); Baek and et al. (2021). Another emerging trend is the use of machine learning force fields, which differ from classical force fields by fitting quantum mechanics calculations through machine learning techniques. These machine-learned molecular mechanics (MM) force fields, such as espaloma-0.3, allow for a more accurate representation of potential energy surfaces without relying on pre-defined atom types Unke *et al.* (2020). While these techniques are powerful, they typically require large datasets to train the models effectively, which can be a limitation Schütt and et al. (2017). The statistical nature of machine learning necessitates access to extensive and diverse data to achieve the desired accuracy and generaliza-

tion capabilities. Nevertheless, the integration of machine learning into structural biology and molecular dynamics represents a significant advancement, promising to enhance the efficiency and accuracy of simulations and predictions in biological systems. Machine learning (ML) and deep learning (DL) have emerged as transformative tools in understanding and optimizing the CRISPR/Cas9 genome-editing mechanism. These computational techniques enable researchers to predict the outcomes of CRISPR interventions, assess off-target effects, and enhance guide RNA (gRNA) design, thereby improving the efficiency and specificity of gene editing. Traditional machine learning algorithms, such as support vector machines (SVM) and random forests, have been employed to predict on-target activity and off-target effects of gRNAs. Recent studies have demonstrated that these models can effectively classify potential cleavage sites based on sequence features, significantly enhancing the understanding of gRNA efficiency and specificity Sherkatghanad *et al.* (2023). For instance, a comparative analysis of ML techniques revealed that random forest models achieved high accuracy in predicting CRISPR-Cas9 cleavage sites, outperforming other methods Abadi *et al.* (2017). The application of deep learning has further advanced predictive capabilities in CRISPR research. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been utilized to analyze complex sequence-activity relationships. These models can capture intricate patterns in the data, leading to improved predictions of gRNA performance Zhang *et al.* (2022). For example, a study highlighted the successful use of artificial neural networks (ANNs) to predict on-target and off-target activities, demonstrating the potential of DL to enhance CRISPR applications Klein and et al. (2022). Additionally, machine learning algorithms have been instrumental in optimizing gRNA design. Tools like Azimuth and Elevation leverage ML to predict the effectiveness of gRNAs in targeting specific DNA sequences. These models assess various features, including protospacer adjacent motifs (PAMs) and gRNA sequences, to recommend optimal gRNAs for desired editing outcomes Listgarten *et al.* (2018). The integration of ML in gRNA design not only improves editing efficiency but also minimizes off-target effects, making CRISPR applications more reliable Klein and et al. (2022). As the field progresses, there is a growing emphasis on the interpretability of machine learning models. Understanding how these models make predictions is crucial for their practical adoption in CRISPR applications. Recent advancements have focused on incorporating attention mechanisms and feature importance analyses to elucidate the decision-making processes of these models, thereby increasing trust in their predictions Sherkatghanad *et al.* (2023)

2.3.1 Machine Learning Methods

Machine learning approaches have been employed in this thesis in Chapter 4 to identify the structural characteristics that aid in classifying the metastable states obtained from Markov State Models (MSMs) for on-target and off-target forms of the SpCas9 and FnCas9 proteins. The workflow is added in the Figure 2.2

Machine Learning modeling to classify heteroduplex states in SpCas9 and FnCas9 systems

MSM was performed to determine the metastable states of the heteroduplex in on-target and off-target forms of the SpCas9 and FnCas9 proteins. Machine learning methods were employed to identify the structural characteristics that aid in classifying the macrostates obtained from MSM, highlighting the essential features contributing to their distinction.

Training and Testing Dataset

The following structural descriptors were used to train the machine learning algorithm (ML). *RNA:DNA hybrid structural properties:* The intra-base pair rotational (propeller, buckle, and opening), translational parameters (shear, stagger, and stretch) of the 20th, 19th, 18th, and 17th base pairs, the number of tight contacts (within 4 Å) between sgRNA-tDNA and the linear interaction energy between both the sgRNA and tDNA strands. These features were computed for the complete trajectory for all on-target and off-target forms of SpCas9 and FnCas9. *Protein structural properties:* The protein-RMSD (backbone atoms), the radius of gyration (ROG), solvent accessible surface area (SASA), the angle between atoms DC-1508, DC-1494, and DA-1527) in SpCas9, and the angle between atoms DC3-1746, DC-1732, and DG3-1765) in FnCas9 and the number of native and non-native contacts between RNA:DNA hybrid and HNH, RuvC and REC1, REC2 and REC3 domains were calculated for the complete trajectory for all on-target and off-target systems of SpCas9 and FnCas9. *Metastable State Determination:* The Markov State Model (MSM) framework was applied to identify the distinct metastable states of the heteroduplex, which were used as labels for the classification task. These labels were also obtained for all the frames present in on-target and off-target forms of both the Cas9s.

Feature Selection Framework

This study used various feature selection and ranking techniques to identify the critical structural determinants responsible for distinguishing metastable states within the SpCas9 and FnCas9 protein systems. Due to the non-linear nature of the relationships between features and the target variable, non-linear machine learning models were employed to investigate their relationships. The dataset for SpCas9 and FnCas9 consisted of 57074 frames and 52080 frames, respectively, where a total of 42 features were extracted for both Cas9 variants. Detailed data distribution has been provided in Supplementary Table S4 (see Supplementary Information). Two primary categories of feature selection methods were utilized: Wrapper and Embedded. Wrapper methods assess feature subsets by feeding them into a model, while Embedded methods identify the most valuable features during the model creation process. Specifically, two Wrapper-based feature selection methods, BorutaSHAP (based on XG Boost), which combines the Boruta feature selection algorithm with Shapley values, and Recursive Feature Elimination (RFE), a technique that iteratively removes features and constructs models with the remaining ones were implemented Gramegna and Giudici (2022); Zeng *et al.* (2009). Additionally, Extra-Tree Classifier, a tree-based embed-

ded feature selection method, was implemented to calculate feature importances while training an ensemble of decision trees. All structural features extracted from the trajectories of on-target and off-target forms of both the Cas9s were input into these three feature selection techniques: BorutaSHAP, Recursive Feature Elimination, and Extra-Tree Classifier Gramegna and Giudici (2022); Zeng *et al.* (2009). The selected features from these methods were used for training and evaluating the machine learning models. Machine learning predictions were performed using three distinct modalities, using: (a) RNA:DNA hybrid structural properties (28) alone, (b) protein structural properties (14) alone, and (c) both RNA:DNA hybrid and protein structural properties (42). Feature selection was independently applied for each approach.

Model Development & Assessment

The dataset was randomly partitioned into training and test sets, maintaining an 80:20 ratio. The input features comprised RNA:DNA hybrid-specific attributes, protein-specific attributes, as well as a combination of both hybrid and protein features. The output variables for these input features were heteroduplex metastable states obtained from MSM for SpCas9 and FnCas9. To ensure the robustness of our feature set, feature selection techniques were employed using Wrapper and Embedded methods. The classification of heteroduplex states was performed with selected features using several different machine learning models (Random Forest, gradient-boosting, ADA-Boost, XG Boost, and Extra-Tree Classifier). The models were trained using ten-fold cross-validation and evaluated using the selected features for all three modalities. Hyperparameter tuning was carried out on selected features and was achieved using Grid Search CV. The models were evaluated using the evaluation metrics Accuracy, Precision, Recall, and F1-score. The accuracy of the model was defined as

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

where TP stands for true positive, FP for false positive, FN for false negative, and TN for true negative. The recall is the measure of how many actual positives the model can capture, while the precision is how many of the predicted positives are correct.

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

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

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

$$F1score = \frac{2 * Recall * Precision}{Recall + Precision}$$

The F1-score is a measure of a model's accuracy, considering both precision and recall, with values ranging from 0 to 1. It was used to compare the model's performance in different modalities. All models evaluated were implemented using the scikit-learn package in Python Kramer and Kramer

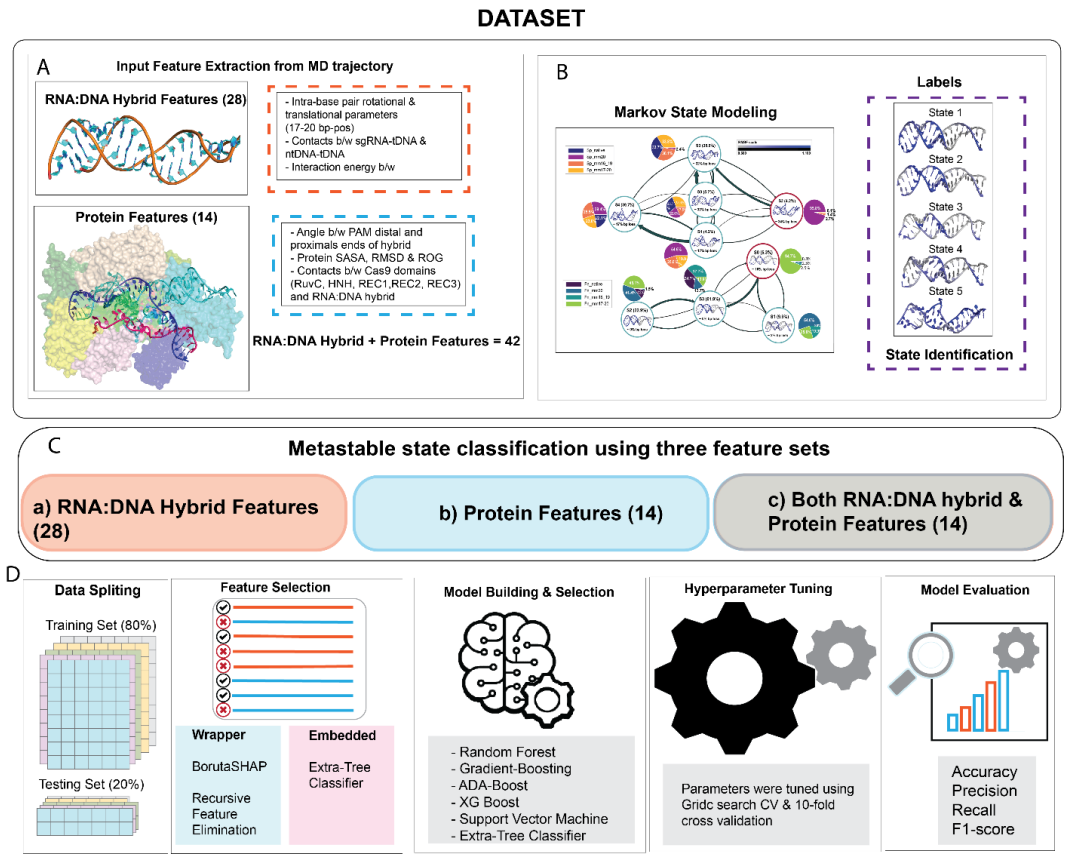


Figure 2.2: The workflow of metastable state classification showing input feature sets, labels and steps followed for prediction. A – Extraction of Input features; B- State identification using MSM. Using states as labels. C- Three features used for prediction of states in both Cas9s (a) RNA:DNA hybrid features, (b) Only Protein features , and (c) both. D. Steps followed for classification

(2016). To identify the optimal machine learning model, the F1 score was used as a selection criterion due to the inherent nature of the predictive task, which involved the classification of outputs with multiclass labels.

2.4 Unraveling CRISPR-Cas9 Dynamics and Mechanisms Using Computational Approaches

Recent developments in MD methodologies, such as coarse-grained models and enhanced sampling techniques, have further expanded the applicability of MD simulations to larger systems and longer timescales. These advancements have proven particularly useful in studying the dynamics of the CRISPR/Cas9 system, where researchers have utilized MD simulations to explore conformational changes and interactions critical for its function. By integrating experimental data with computational simulations, MD has become an indispensable tool in the field of structural biology, enabling the exploration of molecular dynamics in ways that were previously unattainable. The structural information gained also help in designing sgRNA and Cas9 variants with altered

PAM specificity. A work by Maji et al. used SpCas9 binding to PAM to test a large number of small molecules for their ability to interrupt the SpCas9-DNA interaction Maji *et al.* (2019). Post-screening of small molecules they identified BRD0539 as a SpCas9 inhibitor (reversible and cell-permeable inhibitors that disrupt SpCas9-DNA binding) and also found a pharmacophore for SpCas9 inhibition using structure-activity relationships. PAM recognition plays a significant role in genetic manipulation, limiting the target site recognition by Cas9 to a small number of sequences. A structurally designed SpCas9 was built to near-PAMless SpRY in a recent study by Walton et al. Walton *et al.* (2020) which can target almost all PAMs with high editing capability and flexibility, which can be used to expand genome-editing technologies for basic and applied research. Some studies have shown activation of immune response and generation of anti-Cas9 antibodies and T-cells against Cas9 in mouse models and human plasma. This immune response against Cas9 delivery can lead to its decreased efficacy in-vivo posing important safety issues related to its use. A review by Mehta et al. Mehta and Merkel (2020*b*) focuses on the immunogenicity of the Cas9 protein and summarizes potential approaches to overcome this problem of immune recognition. The approaches suggested in this review included structural modification of Cas9 proteins by either by covering up immunogenic epitopes or by using Cas9 orthologs from nonpathogenic bacteria, triggering immune tolerance or immunosuppression or to target immune privileged organs, such as the eye. A recent review by Meaker et al. Meaker *et al.* (2020*a*) talks about the design principles used to engineer diverse functionalities of CRISPR/Cas9 enzyme. The strategies followed for modification of CRISPR/Cas9 systems for genome editing activities, involved mutagenesis of Cas9 domains, reconstruction of the gRNA, integration of additional enzymatic domains to Cas9, and sampling other species for naturally occurring CRISPR variants with better conformational activation, target identification, and DNA cleavage. In their recent paper Zuo and Liu (2020), Zuo and Liu discuss the fundamental allosteric regulation of Cas9 complex assembly and targeting specificity. They also listed a few allosteric modulation measures to mitigate intrinsic Cas9 off-target effects. Their final statement stated that allosteric modulations could be used to create a highly precise and tunable CRISPR/Cas9 system, which would support applications that need both conditional control and high precision. In another work Meaker *et al.* (2020*b*) the authors have tried to increase the specificity of Cas9 cleavage by using mathematical modelling and experiments and came up with this conclusion that the specificity of CRISPR/Cas9 genome editing can be enhanced by coupling Cas9 to artificial inhibitory domains. A review by Chen et al. Chen *et al.* (2019) highlights the development of CRISPR/Cas dependent genome editing system and summarizes different methods feasible for detection of potential off-target effects due to genetic manipulation by the Cas9 enzyme.

Molecular dynamics (MD) simulations have provided valuable insights into the dynamics and mechanisms of CRISPR-Cas9, contributing to the understanding of its function, catalysis, allostery, and specificity. By combining classical and coarse-grained approaches, researchers have been able to access a wide range of spatial and temporal scales, shedding light on the complex conformational changes and interactions that underlie the CRISPR genome editing process. The first classical molecular dynamic (MD) simulation on the CRISPR system was performed in 2015 by Estarellas et al. Estarellas *et al.* (2015) using an RNA/Csy4 complex without DNA, in explicit-

solvent, in order to understand its basic structure and dynamics. Due to a few limitations in the simulation techniques, like time constraints, the accuracy of the force field, and the resolution of available crystals, the results were not reliable enough. In 2016, Wenjun Zheng and his group Kleinstiver *et al.* (2016*b*) tried to overcome the time constraints by performing coarse-grained MD simulations of a ternary CRISPR/Cas9 complex consisting of RNA, DNA, and protein. The research identified a few key protein regions that can be altered to improve Cas9 specificity, as well as new dynamic insights into allosterically induced structural rearrangements of the protein's proximal and distal domains in response to sgRNA and DNA binding.

2.4.1 Conformational Rearrangement of Cas9 endonuclease on Nucleic Acid Binding

In 2017, a set of four all atom MD simulations was performed by Palermo et. al. that focussed on the importance of the presence of ntDNA for HNH domain activation and tDNA strand cleavagePalermo *et al.* (2017). Molecular dynamic simulations were carried out for four different systems free-Cas9, Cas9 bound to RNA (binary complex), Cas9 bound to RNA and DNA in an intermediate state (ternary complex, PDB ID: 4UN3) Anders *et al.* (2014*a*) and Cas9 ternary complex in a pre-catalytic state from PDB ID: 5F9R Jiang *et al.* (2016). This paper discusses Cas9's identification and binding, of RNA, which has previously received little attention. Two conformational changes take place during the CRISPR/Cas9 classification, one when the Cas9 converts from apo to RNA bound form (involving the REC lobe of Cas9), suggesting RNA binding has taken place, second when HNH domain transforms Cas9 in catalytically active conformation to leave the DNA. On the RNA binding the REC lobe opened up, exposing the arginine helix into the solvent, forming a positively charged cavity to accommodate the RNA. All the three REC lobes moved in different directions. The experimental and computational studies were in a line indicating that on recognition of the nucleic acid the protein domains move jointly in different directions. The results from the MD simulations revealed that in the absence of ntDNA, the scissile phosphate of tDNA is away from His840 of HNH domain (25 Å), unable to cleave the tDNA, while in its presence it is only 15 Å apart from His840 of HNH domain. The same work reported that hinge regions LI (residues 765-780) and LII (residues 906-918-K913) that links the RuvC and HNH domains form interactions with ntDNA that stays throughout the simulation time, hence stabilizing His840 at a smaller distance of 15 Å from the scissile phosphate of the tDNA strand. This positioning of HNH by RuvC and ntDNA interactions, is termed as 'crosstalk' between HNH and RuvC domain suggesting that ntDNA plays a pivotol role in Cas9's activation. The conformational plasticity obtained by the HNH domain allows it to reposition itself, allowing the RuvC domain to bind with ntDNA, according to comparative MD simulations of CRISPR/Cas9 ternary complexes in pre-catalytic (PDB ID: 5F9R) and post-catalytic (PDB ID: 4UN3) states. In this study, the active and inactive states of HNH domains are shown in a distinguished manner using GaMD simulations. The results from the tMD simulation speculated that the flexibility of HNH domain, via which it undergoes a major structural transition during its activation, causes unwinding of dsDNA and

formation of R-loop structure. These studies implicate that the conformational changes required in RNA recognition, DNA unwinding and cleavage can be explained in great detail using Molecular Dynamics, Targeted Molecular Dynamics, and Gaussian Accelerated Molecular Dynamics.

In 2016, Jiang et al., used cryo-electron microscopy (cryo-EM) technique to crystalize the structure of SpCas9 in a complex with sgRNA and dsDNA containing the PAM sequence to determine the structural basis for R-loop formation and DNA cleavage by the two catalytic Cas9 domains. The authors observed that due to ntDNA binding the REC and NUC lobes move toward the central nucleic acid–binding channel making the Cas9 structure more compact. The most prominent change due to dsDNA binding occurred in the HNH nuclease domain, which rotated by $\sim$180° and translate by $\sim$ 20Å towards the RNA-DNA hetero-duplex leading to the positioning of His840 adjacent to scissile phosphate. They also observed that dsDNA substrate binding to Cas9 led to improved stability as compared to its association with a single target DNA strand or a PAM-containing partial duplex. The two hinge regions connecting the HNH domain with the RuvC domain—L1 (residues 765 to 780) and L2 (residues 906 to 918) underdo structural rearrangements stabilizing the R-loop and thus aid in HNH domain repositioning Jiang *et al.* (2016). However, in the structure given by Jiang et al., the distance between the Cα atom of catalytic residue His840 and the scissile phosphate was still more than 19 Å. In 2017, Huai et al. resolved a 5.2 Å cryo-EM structure of SpCas9 in a complex with sgRNA and target DNA in the DNA cleavage-activating state. They captured a distinct SpCas9 state in which the HNH active site is the closest to the scissile bond of the target DNA by contacting the PI and REC1 domains Huai *et al.* (2017). They also observed that the HNH domain exists in three conformational states, the distance between the Cα atom of residue His840 and the scissile phosphate reduced from $\sim$32 Å (state 1) and $\sim$19 Å (state 2) to $\sim$10 Å (state 3).

Using single molecular FRET, Osuka and his team tried to reveal the flexible inter-domain movements that allow Cas9 to adopt transient conformations. Upon sgRNA binding, a decrease in the flexibility between the REC and NUC lobes was observed. In contrast, the target DNA binding increased the percentage of fluctuating molecules suggesting the increased flexibility between the REC and NUC lobes in the Cas9-sgRNA-DNA ternary complex suggesting that sgRNA binding stabilizes Cas9, while the binding of DNA and Mg2+ enhances inter-domain fluctuations Osuka *et al.* (2018). According to a Cryo-EM study performed by Zhu et al., upon sgRNA binding to apo-Cas9, substantial changes in the Cas9 protein domains were observed. Following Mg2+ and DNA binding to the binary complex, an intermediate state I or "checkpoint" conformation is formed (Figure 2.3A(i)) which is a proof-reading state that determines if sufficient base pairing has been created between the sgRNA and TS. In this state, the HNH nuclease domain is positioned far away from the DNA. The newly displaced NTS travels parallel to the RNA-DNA hybrid within the central channel after RNA-DNA hybridization, while REC3 lobe screens for correct base-pairing between TS and sgRNA at the PAM-distal end. The HNH domain swings dramatically $\sim$34 Åand the REC2 recognition domain becomes disordered during the transition to the post-catalytic state. The cleaved substrate coupled to the catalytically competent HNH active site is captured in the post-catalytic state. The HNH domain is disordered in the product state, where the REC2 reverts

to its pre-catalytic conformation, and further interactions between REC3 and RuvC and nucleic acids are produced, as can be seen in Figure 2.3A (ii-iii) Zhu *et al.* (2019).

Recently, the structural changes associated with R-loop formation and conformational changes associated with it were elucidated using the Cryo-EM technique Donohoue *et al.* (2021). In the gRNA bound form of SpCas9, the REC2 and REC3 domains obstruct the primary DNA binding channel. The REC2 domain is shifted to produce a binding cleft to accommodate the PAM-distal DNA duplex after PAM recognition and the first 5-nt base pairing with the guide RNA seed sequence. The REC3 domain is further displaced and the central binding channel is entirely opened by the formation of an 8-bp hetero-duplex, while the PAM-distal duplex remains in the REC2/3 cavity. The gRNA-TS hetero-duplex and the PAM-distal duplex are placed into the central binding channel by extending the R-loop to a 10-bp hetero-duplex, which is accompanied by the creation of electrostatic contacts between the REC2 and REC3 domains. Base pairing past the seed region results in undocking of the HNH domain from the RuvC and PI domain interface, and results in its re-positioning towards target hetero-duplex into the checkpoint state. R-loop formation past 18 base pairs induces REC2 domain displacement from the binding channel and rotation of the HNH domain active site towards the TS cleavage site, while simultaneously positioning the NTS in the RuvC domain active site. Further, this study highlighted the importance of maintaining guide-target complementarity and proper hetero-duplex geometry. Using a single molecule fluorescence technique, the HNH nuclease domain motions from DNA binding to post-DNA cleavage was studied in real time Wang *et al.* (2021). They revealed a ntDNA cleavage competent state and concluded that the HNH domain is highly flexible post-TS cleavage, explaining a recent observation that the HNH domain was not visible in a postcatalytic cryo-EM structure. They also discovered that the 3' flap generated after NTS cleavage gets digested by a ssDNA specific exonuclease, suggesting Cas9 exposes the 3' flap for potential interaction with DNA repair machinery.

2.4.2 Allosteric Mechanism of Cas9 protein in PAM recognition

The first step in the CRISPR/Cas system is recognition of the PAM sequence, the sgRNA guides Cas9 to the PAM sequence where the Cas9 binds it and cleaves the DNA at a site away from the sequence showing an allosteric effect. Two arginine residues (R1333 and R1335) mediate the recognition of the '5-TGG-3' PAM sequence in CRISPR/Cas9 (Figure 2.4A. A better understanding of this allosteric mechanism via which the PAM sequence coordinates with catalytic domains of Cas9 (HNH & RuvC) can guide in regulating the functionality of CRISPR/Cas9s Sternberg *et al.* (2015); Palermo *et al.* (2017). In order to understand the allosteric mechanism of Cas9, Palermo et al(2017), performed 13 microseconds long molecular dynamics simulations using two systems Cas9 with guide RNA and tDNA with PAM and without PAM (5'-TGG-3'), under suitable simulation conditions for DNA/RNA endonucleases. The results stated that PAM serves as an allosteric effector, stimulating the interdependent conformational dynamics of the HNH and RuvC domains, which are responsible for bringing double strands breaks in DNA similar to what was found in FRET experiments by Sternberg et. al. Sternberg *et al.* (2015) Cas9 adopts different

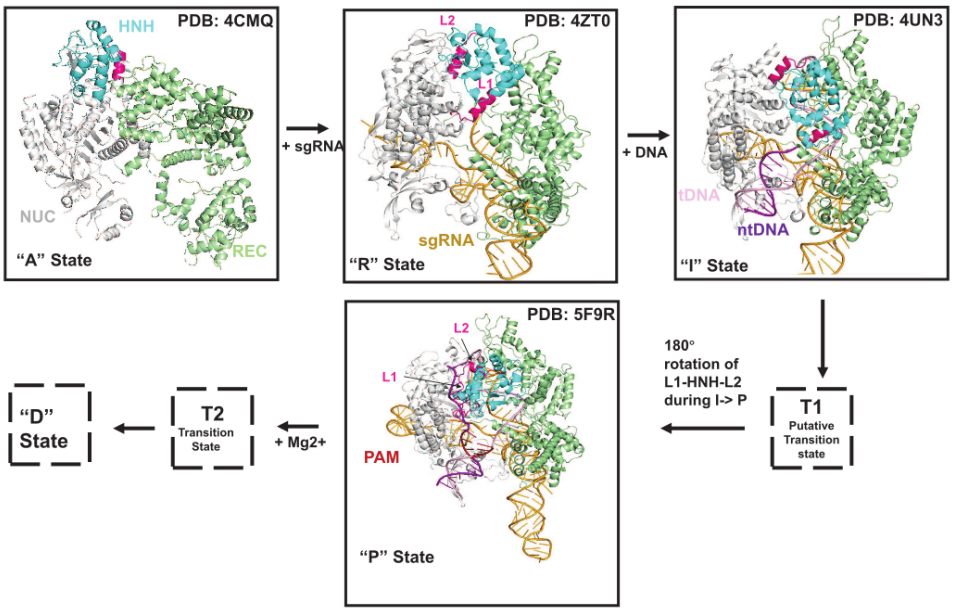


Figure 2.3: A. Pre-catalytic to catalytic conformational transitions in the HNH domain. In cyan and magenta, the HNH domain and flanking linker regions (L1 and L2) are shown, respectively. The Protospacer Adjacent Motif is shown in dark red(Palermo (2019*a*)).

conformations when interacting with PAM and PAMless DNA, demonstrating an "open-to-close" conformational transition that correlates to the conformational transition necessary for nucleic acid association. In the presence of PAM, the HNH and RuvC domains are strongly interconnected and also HNH has a strong correlation with the REC domain, which is in agreement with the experimental evidence that HNH dynamics regulate Cas9's nuclease behavior. Palermo et. al. in their recent work highlighted the residues involved in the allosteric pathway (Figure 2.4D. The charged amino acid residues, Gln771, Glu584, Lys775 and Arg905 are essential nodes of the dynamic network in Cas9 with PAM suggesting that electrostatic interactions play a significant role in allosteric signaling. Interactions among these residues connect HNH to RuvC and the REC lobe via the linkers LI (765-780) and LII (906-918), which effectively function as "allosteric transducers" (Figure 2.4C) Jinek *et al.* (2012*a*); Sternberg *et al.* (2015). The flow of information to catalytic residues in the HNH and RuvC domains takes place via Lys772 and Thr770 residues of the L1 loop. East et al. examined the allosteric process of information transmission through the HNH domain using a blend of NMR experiments and MD simulation East *et al.* (2020). The HNH domain's full allosteric pathway extends from the area where the RuvC domain meets the HNH domain to the RECII region. The recognition lobe and nuclease domains (HNH and RuvC) transmit the DNA binding signal for coordinated cleavage of the t-DNA and nt-DNA strands. They discovered the existence of a millisecond timescale dynamic pathway using a novel engineered HNH construct which was conserved in the full-length Cas9, confirming the existence of a communication channel between REC2–HNH–RuvC that preserves the conformation and versatility of full-length Cas9. The functional role of the allosteric pathway was identified by introducing alanine mutations into all the residues involved in the pathway (Gly792, Gln794, Lys797, Glu798, Leu816, Gln817, Asn818, Arg820, Val838, His840, Ile841, Lys855, Figure 2.4B) and computing

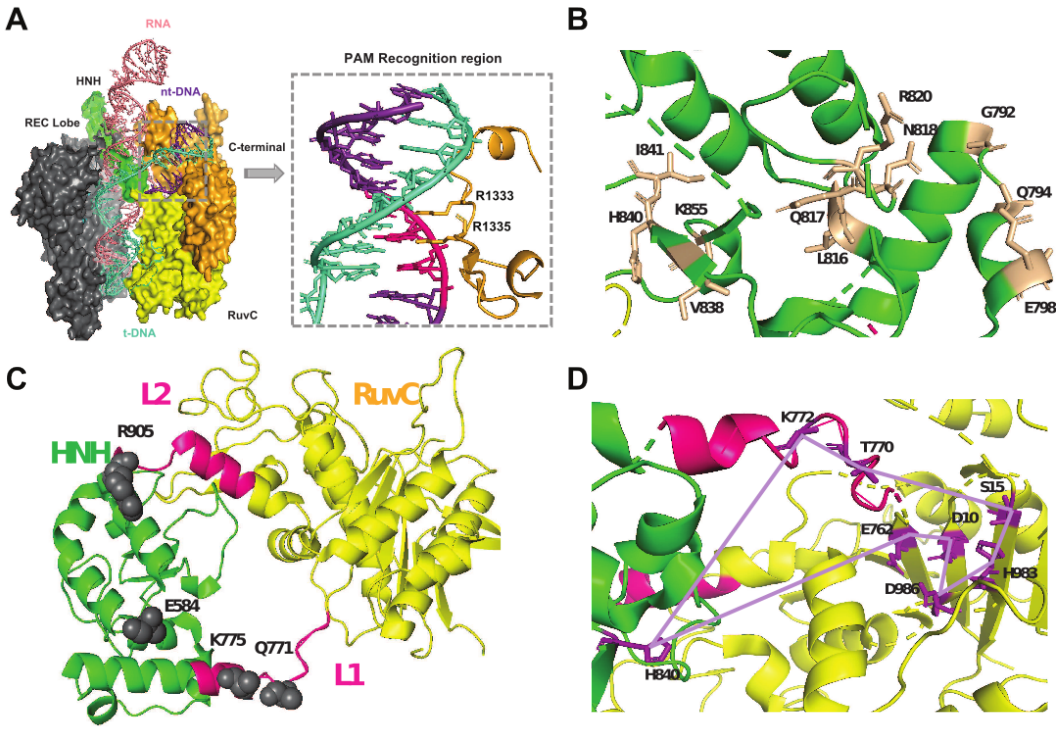


Figure 2.4: A. SpCas9 is bound to a gRNA and a tDNA (4UN3.pdb) architecture with a close-up view of the PAM recognition region. B. Residues are shown in sticks representation in which alanine mutations were introduced in East et al. C. At the essential edges of the allosteric pathway (Tyr771–Lys775 and Glu584–Arg905) (grey spheres) connection between the L1 (residues 765–780) and L2 (residues 906–918) loops (magenta) linking HNH (776-905) (green) and RuvC (red) domains is shown. D. "Shortest pathways" linking pairs of catalytic residues of the RuvC and HNH domains of the Cas9 with PAM.

the communication between amino acids. It was observed that the substitution of key allosteric residues affects the selectivity of CRISPR/Cas9. The crosstalk between the RuvC and HNH helps Cas9 to cleave the target and non-target DNA strands in a concerted manner, as seen in this analysis. PAM is needed for Cas9 to recognize DNA and to activate the two catalytic domains. In a recent paper, Palermo *et al.* (2019) a mathematical approach was used to research allosteric communications in the CRISPR/Cas9 complex. Shannon's entropy was used to distinguish pairs of atoms that showed interdependent motions during the simulations. After the atoms have been identified, a pairwise correlation coefficient is calculated to determine how much one atom's location limits the position of another and vice versa. PAM, the small sequence of DNA that initiates recognition and binding, turns out to be an allosteric effector that promotes contact between RuvC and HNH, the two catalytic domains.

2.4.3 Mechanism of target and cleavage by Cas9

Experimental techniques like, ssDNA curtain assay, atomic force microscopy (AFM), and single molecule resolution have revealed that Cas9 finds its target by direct 3D collision. The study by

H. Nishimasu et al. Nishimasu *et al.* (2014*a*) gives some key steps involved in the Cas9 mediated cleavage process. The Cas9 protein when unbound to single gRNA(sgRNA) present in its inactive state. When the sgRNA binds to it and forms Cas9-sgRNA complex, Cas9 undergoes a conformational change, reordering PAM (Protospacer adjacent motif) Interaction domain allowing the scanning of target DNA. The complex then looks for matching PAM motifs in the non-target DNA and crRNA sequence in the target DNA strand. Once the match is found, it binds to it, resulting in another conformational change in the REC lobe of Cas9. The crRNA interacts with the target DNA forming a channel for RNA:DNA hetero-duplex which brings about a domain rearrangement in the domain of Nuclease lobe. The HNH domain allosterically regulates the RuvC-like domain by shifting into a catalytically active position. These two domains cleave the double stranded DNA, by bringing double stranded breaks into it. The findings indicate firstly, that the specificity of the Cas9 to recognize the target DNA and cleave it is conferred by crRNA of single guide RNA, secondly, hetero-duplex(RNA-DNA) formation begins at the PAM motif and proceeds to unwind the target DNA. Palermo et al. in 2017 reported the first model of fully catalytic CRISPR/Cas9 protein Palermo *et al.* (2017). By MD simulations, Zuo and Liu Zuo and Liu (2016) discovered the catalytically competent state of the RuvC domain in 2016, although the HNH conformations remained mysterious. They extended their study and reported the conformational activation of the HNH domain from pre-catalytic to catalytic state in 2017 Zuo and Liu (2017); Yoon *et al.* (2019) by employing biased targeted-MD (tMD) and unbiased ensemble conventional MD (cMDens). The crystal structures of SpCas9 at different binding stages Jiang *et al.* (2016); Fonfara *et al.* (2014) were used to understand the activation and structural rearrangement of Cas9-HNH domains (Figure 2.3A). Zuo and Liu obtained a catalytically active state of the Cas9 HNH domain prepared for cutting the tDNA strand. This research also revealed that Mg2+ ions are necessary for the development and maintenance of the HNH domain's catalytic state, which is consistent with Jinek et al. Palermo *et al.* (2016*a*) and Dagdas et al. experimental findings Chen *et al.* (2017*b*). In a recent work Zhu *et al.* (2019), Zhu, Clarke, Puppala using cryo EM technology, provided snapshots of the Cas9 enzyme in action as it performs concerted cleavage of a target DNA. The author discovered three distinct states of Cas9 after imaging a single cleavage reaction. State -I, represents a checkpoint state prior to HNH-RuvC mediated cleavage, whereas states II and III reflect structural reorganization occurring after catalysis and conversion to a subsequent product-bound intermediate (Figure 2.5A). On the basis of 16s unbiased MD simulations performed with Anton-2, Palermo et al. (2018) revealed that following DNA cleavage, the structural transformation in the HNH domain is followed by a dramatic structural remodelling of the REC lobe Palermo *et al.* (2018). The HNH domain is regulated and locked at the cleavage site by this lobe, resulting in a catalytically active CRISPR/Cas9 complex. Using the empirical valence-bond (EVB) process, Yoon et al. Zuo and Liu (2017); Yoon *et al.* (2019) sought to determine the catalytic mechanism and energetics for Cas9 activation. His group utilized endonuclease VII, a Cas9-like protein, to discover that the catalytic state of the HNH domain contains a structure in which the K848 residue is closer to the scissile phosphate of the tDNA strand. The Palermo group performed a QM/MM simulation at DFT level which revealed in 2019 a new configuration for the active site. (Figure 2.5B) Palermo (2019*a*). Their results indicate that phosphodiester bond cleavage by two-metal

aided catalysis involves only the scissile phosphate's pro-Sp oxygen to coordinate with the two Mg2+ ions, allowing for an in-line SN2 nucleophilic attack, and their findings are consistent with the catalysis' requirements. One water molecule interacts with H983 in close proximity to the scissile phosphate. As required for two-metal aided catalysis, the oxygen pro-S (Sp) coordinates the two metal ions jointly. Results in this study raise several unanswered questions like how the two metal ions (Mg2+) rearrange themselves to achieve the catalysis, the origin of the nucleophile that initiates the chemical process, the role of His983 in catalytic process, as well as the kind of catalysis that results in phosphodiester bond breaking, are unknown Casalino *et al.* (2020).

2.4.4 Mechanistic insights into off-target effects

The issue of off-target effects is a significant concern in the application of CRISPR-Cas9 genome-editing technology, as it refers to the unintended cleavage of DNA sequences that do not match the guide RNA, potentially leading to undesirable phenotypes. Molecular dynamics (MD) simulations have been utilized to investigate this phenomenon by introducing base pair mismatches within the RNA:DNA hybrid of the CRISPR-Cas9 complex. Inspired by single-molecule experiments, these studies have shown that the binding and catalytic activity of Cas9 are highly sensitive to the positioning of mismatches. Specifically, mismatches located within nine base pairs of the protospacer adjacent motif (PAM) can dramatically affect DNA binding and cleavage, while mismatches further away may still allow stable binding but compromise catalysis. Studying the dynamics of Cas9 variants using MD simulation have also aided in understanding off-target effects. Kleinstiver and his coauthors Kleinstiver *et al.* (2016*b*) performed an experiment that shows the ability to decrease nearly all genome-wide off-target effects. Based on Kleinstiver's work, Zheng's group Zheng *et al.* (2018) chose four mutants: Asn497A, Arg661A, Gln695A, and Gln926A (referred to as nrqqCas9), and provided the crucial atomistic justification for the decrease of off-target effects accompanied by a decrease in on-target cleavage efficiency of nrqqCas9. The nrqqCas9 systems follow a partially open conformation with lesser interactions between the HNH and RuvC domains, resulting in lower on-target and off-target DNA cleavage performance. Slaymaker and his group performed an experiment Slaymaker *et al.* (2016*a*), that attractive electrostatic interactions between positively charged residues at the interface-region of the HNH and RuvC domains and the negatively charged ntDNA strand helps in keeping the ntDNA unwounded (Figure 2.5C), thereby restricting the hybridization of the DNA duplex and keeping the system ready for cleavage. Based on this observation, Ray et al. validated this model by a combination of MD simulations (1.5μs, PDB code: 4UN3 with ntDNA) and electron paramagnetic (EPR) measurements of spin-labeled CRISPR/Cas9 ternary complexes, with nitroxide spin labels on the DNA backbone. Neutralizing these residues at the interface could weaken the binding of ntDNA strand to the protein leading to rehydridization between the tDNA and ntDNA strands, causing "R-loop collapse" Ray and Di Felice (2020). A classical all-atom MD simulations of ternary complexes of wild-type SpCas9 and three of its mutants: Lys855Ala, His982Ala and the combination Lys855Ala & His982Ala was executed for over a cumulative time scale of 10s. This study's conclusion emphasises the

important structural effect of mutation on Cas9 specificity. In complexes with Cas9 mutants, it was discovered that the “unwound” portion of the non-target DNA strand has more versatility and seeks to shift away from the HNH/RuvC interface, where it is otherwise stabilised by electrostatic interactions in the wild-type complex. Palermo and colleagues published a paper in 2017 Palermo *et al.* (2017) that used GaMD simulations to decode the mechanism of off-target binding at the molecular level. They characterised the molecular determinants of off-target binding at the atomic level using these MD simulations, providing crucial insights into the mechanism of off-target effects in CRISPR/Cas9. Their study revealed that DNA containing one to three mismatches near the PAM-distal end can maintain sufficient flexibility in the HNH domain to adopt an active catalytic conformation. However, when four mismatches are present, the HNH domain often remains in an inactive state. MD simulations have revealed that these four mismatches lead to an extended opening of the minor groove in the RNA:DNA hybrid, which enables novel interactions with the L2 loop of the HNH domain. This interaction effectively "locks" the HNH domain in an inactive configuration, reducing its ability to undergo the necessary conformational changes for DNA cleavage. In contrast, internal mismatches within the RNA:DNA structure do not significantly disrupt the interaction with the protein. Recently, in the presence of mismatches, using kinetics guided Cryo-EM, the different stages of Cas9 DNA cleavage were determinedBravo *et al.* (2022*b*). The study observed that mismatches present distal to the protospacer adjacent motif are stabilized by reorganization of a loop in the RuvC domain. The formation of R-loop activates the Cas9 protein and prepares it for catalysis. The gRNA–TS duplex acquires a distinct linear conformation during R-loop propagation. REC3 domain captures the PAM-distal end of the linear duplex after the R-loop is completed. The presence of mismatches in the PAM distal end, inhibit REC3 binding, thereby, preventing Cas9 activation in the following stages. Following the creation of the kinked R-loop conformation, the L1 and L2 linkers use the gRNA–TS duplex as a scaffold to place the HNH domain at the target strand’s scissile phosphate and the NTS in the RuvC site , allowing Cas9 to make a double-strand break. Mutations in the RuvC loop (equivalent to SuperFi-Cas9), according to this hypothesis, prevent the development of the kinked conformation and subsequent cleavage of the gRNA–TS complex. These findings based on MD simulations have been supported by various X-ray crystal structures that capture the binding of Cas9 to off-target substrates. Overall, MD simulations, single-molecule experiments, and structural analyses have collectively advanced our understanding of off-target binding in the CRISPR-Cas9 system.

2.5 Conclusions

CRISPR-based technologies have immense potential by being an addition to the genome editing toolbox for developing biotherapies that will enhance patient outcomes in the future Ghosh *et al.* (2019). The findings related to CRISPR/Cas9 system discussed in this review shall be helpful in revealing any structural, physicochemical, and functional details, which could then be utilized in the structural characterization of this system and deciphering the dynamics at the atomic level. Such insights could be very crucial for optimizing and further expanding the applicability and

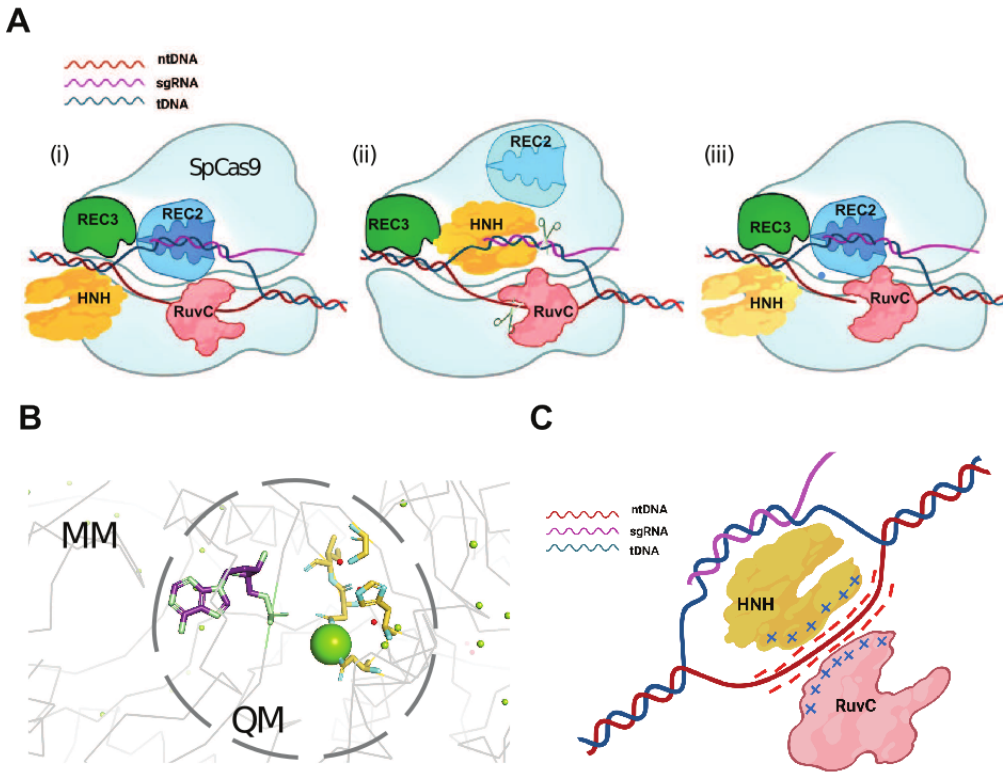


Figure 2.5: A. Snapshots of Cas9 in action-(i) Checkpoint state: Loading of sgRNA into SpCas9. At this state, the HNH domain remains disordered while the enzyme investigates complementarity between the sgRNA and tDNA. (ii) Post catalysis conformation: The REC2 domain dissociates from the tDNA strand and the HNH domain undergoes a conformational rearrangement placing the catalytic residues in line with the phosphodiester backbone. (iii) Product State: SpCas9 back to checkpoint state with ordered REC2 domain and tDNA and ntDNA held by the REC2-REC3 domains and RuvC domain, respectively. These interactions are important for the stability of Cas9 bound to tDNA. B. QM/MM partitioning of the RuvC active site in CRISPR/Cas9. The QM part includes the Mg2+ ions surrounded by residues coordinating the metal ions. C. System's multiple competitive interactions are depicted in this diagram: Positively charged residues (blue color) at the boundary between HNH (yellow) and RuvC (pink) domain have electrostatic interactions with the negatively charged ntDNA (dark red) strand that keeps the ntDNA unwounded.

design of smaller, specific, and effective CRISPR/Cas systems. Future research concentrating on gaining a deeper knowledge of the intracellular processes behind CRISPR-Cas9-mediated gene editing shall empower more accurate clinical applications.

CHAPTER 3

Comparative Structural and Dynamics Study of Free and gRNA-bound FnCas9 and SpCas9 Proteins

3.1 Introduction

CRISPR/Cas (clustered regularly interspaced short palindromic repeat-CRISPR associated) systems confer viral resistance by recognizing and cleaving invading nucleic acids and have been found in the genomes of most archaea, and nearly half of bacteria Brouns *et al.* (2008*b*); Barrangou *et al.* (2007). The CRISPR/Cas immune response involves three stages - spacer acquisition/adaptation, crRNA expression-processing, and interference. The CRISPR/Cas system consists of a DNA cleaving enzyme, CRISPR-associated endonuclease (Cas protein), and its associated gRNA components (preceding the Cas operon), crRNA (CRISPR RNA), and tracrRNA (transactivating gRNA, non-coding RNA links Cas9 with crRNA) which recognize the DNA cleaving region and initiate a double-stranded break.

The most often employed DNA cleaving enzyme from the type-II CRISPR/Cas system is SpCas9, which is derived from the bacteria *Streptococcus pyogenes*. SpCas9 enzyme is 1368aa long (4.1 kb) and adopts a bilobed architecture containing a nuclease lobe (NUC) and alpha-helical lobe or recognition (REC) lobe Chen *et al.* (2014). The nuclease lobe (residues 51-55 and 719-1368) comprise HNH (residues 780-906) and the RuvC-like domain residues (residues 1-55, 719-765, and 919- 1099), Topoisomerase domain (residues 1099-1200) and C-terminal domain (residues 1201-1368). The HNH and RuvC domains help in DNA cleavage by bringing double-strand breaks into the DNA. The recognition lobe contains a 'bridge-helix' (residues 59-76) which connects the NUC and REC lobes with positively charged arginine residues Nishimasu *et al.* (2014*b*); Josephs *et al.* (2015*b*). Despite the rising use of the CRISPR/Cas9 technology in genome engineering, there are significant constraints that prevent it from being widely used. This includes Cas9's large size (which makes delivery difficult), recognition of a specific PAM (which limits its efficiency), and the introduction of random off-target mutations in target gene sequences. The CRISPR/Cas system's site-specific cleavage fails 15% of the time Zhang *et al.* (2015*a*); Cho *et al.* (2014*b*); Wu *et al.* (2014), resulting in unintended genetic alterations, and off-target effects. This limitation of Spcas9 led to the creation of novel Cas variants (e.g., Hifi Cas9, eSpCas9(1.1), SpCas9-HF1, SpCas9-NG, xCas9, Cpf1) with better efficiency Manghwar *et al.* (2019); Casini *et al.* (2018*a*); Lee *et al.* (2018*a*); Vakulskas *et al.* (2018*a*). The varying specificity and sensitivity of Cas variants emphasize the relevance of different Cas9 domains.

In a work by Acharya et al., an ortholog of Cas9 from *Francisella novicida* (FnCas9), was shown to have very low non-specific editing compared to SpCas9 Acharya *et al.* (2019*a*). FnCas9

consists of 1,629 amino acids and is significantly larger than other Cas9 orthologs such as SpCas9 (1,368 amino acids) and SaCas9 (1,053 amino acids). FnCas9 comprises seven domains — the REC1–3, RuvC, HNH, WED (Wedge), and PI (PAM interacting) domains. The structural comparison of FnCas9 with SpCas9 revealed unanticipated conserved and divergent features of inter-domain interaction Hirano *et al.* (2016). In contrast to *Streptococcus pyogenes* Cas9 (SpCas9), which has shown variable levels of off-targeting due to tolerance of mismatches predominantly in the "non-seed" region of the gRNA, FnCas9 has been reported to have strong intrinsic specificity for its targets by tolerating only a single mismatch in the gRNA at the $5'$ position of the PAM distal end Hirano *et al.* (2016). Acharya et al. reported the specificity of FnCas9 is determined by its minimal binding affinity with DNA in the presence of mismatches, the staggered cleavage at the cleavage site,and higher homology-directed repair rates Acharya *et al.* (2019*a*). They have also suggested the basis behind the specificity of FnCas9 could be due to the interaction of its substrate with expanded REC3 and REC2 domains which are structurally different from those in SpCas9 and SaCas9 Hirano *et al.* (2016). They also speculated that FnCas9's high electrostatic potential and interactions with several bases on PAM distal and proximal ends of the substrate are necessary for FnCas9 to achieve its cleavage competent state.

Extensive structural and mechanistic studies involving atomistic simulations, targeted-MD, and accelerated-MD provided interesting insights into the understanding of the recognition mechanism of SpCas9 domains, the mode-of-action, and the kinetics behind SpCas9 mediated DNA cleavage. In 2016, Palermo et al. performed all-atom molecular dynamic simulations of Apo-SpCas9, SpCas9-gRNA (binary complex) state, SpCas9-gRNA-tDNA bound to incomplete ntDNA and complete ntDNA state to reveal the conformational reorganization of Cas9 associated with nucleic-acid binding. They identified that the catalytic residue H840 was at a distance of 25 Å from the scissile phosphate of the tDNA in the absence of ntDNA, whereas in the presence of ntDNA, H840 stabilizes at a smaller distance of 15 Å from the scissile phosphate of the tDNA, suggesting that ntDNA binding is crucial for the formation of a catalytically competent Cas9 Palermo *et al.* (2016*b*). Gaussian Accelerated MD ($\sim$ 16 μs) on the pre-catalytic state of SpCas9 (PDB ID: 5F9R) showed how upon DNA-binding, the conformational dynamics and flexibility of HNH domain might facilitate the unwinding of dsDNA and formation of R-loop structure triggering the formation of active state Jiang *et al.* (2015); Palermo *et al.* (2017). Ab-initio MD study of the pre-catalytic state and active-state of SpCas9 was performed to disclose the two-metal-dependent mechanism of phosphodiester bond cleavage by RuvC Active site of CRISPR/Cas9 Palermo (2019*b*).

While much has been studied for the SpCas9 Palermo *et al.* (2016*b*, 2017); Palermo (2019*b*), the cleavage and recognition mechanism of FnCas9 and the factors responsible for its higher specificity are yet to be explored. Questions pertaining to the molecular inter-play for the interactions and a comparison between the two orthologs can provide us insights into the elusive future of designing customized Cas9 molecules. To address these questions, crystal structures of gRNA-free and gRNA-bound forms of SpCas9 and FnCas9 complexes were simulated for 1 μs and analyzed. The principal component analysis, free energy landscape, and clustering methods were used to

investigate the slow motions of Cas9 proteins and conformational changes due to the binding of gRNA. The relative domain movement was studied by distance plots between the center of mass of each domain pair. The crucial interactions and key residues involved in the binding of gRNA with Cas9 protein were examined at the Cas9 protein-RNA interfaces by binding free energy calculations.

Materials and Methods

Estimation of evolutionary conservation of Cas9 domains

The amino-acid positions which evolve slowly are commonly conserved sites that are important for protein structure and function. Proteins alter their secondary structure due to a change of surroundings or as a result of interaction with other proteins, but these conserved residues preserve their structures in all the conformations of the protein and therefore contribute less to the structural flexibility of the protein Chen *et al.* (2007). The ConSurf server Ashkenazy *et al.* (2016) was used to estimate the level of conservation of amino acid positions in different domains of the SpCas9 and FnCas9 protein based on phylogenetic relationships between their homologous sequences. The protein structure of SpCas9 (PDB ID: 4CMQ) and FnCas9 (PDB ID: 5B2Q) were submitted to the ConSurf server, which calculated the conservation scores partitioned into the discrete scale of nine bins. The position with bins $>$ 8 scores is considered conserved sites, while the amino-acid positions with bin <= 2 are considered variable sites. Applying these cut-offs for indicating conserved and variable residues, the percentages of conserved and variable residues were calculated for HNH, RuvC, BH, REC1, REC2, and REC3 domains in both the proteins. The proportion of charged (H, K, R, D, E), uncharged (C, S, T, Y, Q), and non-polar (A, G, I, L, M, W, F, P, V) residues in conservation was also calculated.

System setup and MD Simulations

MD simulations have been performed using four model systems of Cas9, which were the Apo form of SpCas9 and in complex with gRNA, and Apo form of FnCas9 and in complex with gRNA. These model systems have been prepared using the crystallographic coordinates of *Streptococcus pyogenes* Apo-SpCas9 (PDB ID: 4CMQ) Jinek *et al.* (2014*a*), SpCas9+RNA (PDB ID: 4ZT0) Jiang *et al.* (2015), *Francsella novicida* Apo-FnCas9 (PDB ID: 5B2Q w/o nucleic acids) Hirano *et al.* (2016) and FnCas9+RNA (5B2Q, DNA strand removed), solved at 3.09 Å, 2.90 Å, and 1.70 Å resolution, respectively. The four independent MD simulations were performed using GROMACS 2020 software package Abraham *et al.* (2015). AMBERff99SB and AMBER94 forcefields were used for protein, and nucleic acid parameterization Guy *et al.* (2012). The system was neutralized and solvated by explicit water molecules, which were modeled by the TIP3P) parameter set Jorgensen *et al.* (1983) in a dodecahedron box. Particle mesh Ewald was used to treat long-range

electrostatic interactions Darden *et al.* (1993). The pressure and temperature were controlled with the Parrinello-Rahman coupling algorithm, and Langevin thermostat Hatano and Sasa (2001). The energy minimization of the whole system was done using the steepest descent algorithm until the maximum force was less than 1000 kJ $mol^{-}1nm^{-}1$ to remove any steric clashes. The system was heated to 300 K and equilibrated for 100 ps in both Isothermal-Isobaric and canonical ensembles. The positions of Cas9 protein and gRNA were restrained with a harmonic constant of 0.1 kcal/mol $Å^{-}2$ to avoid improper geometry. A non-restraint MD simulation was performed for 1000 ns at constant 1 atm pressure. The cut-off value for nonbonded interactions was set to 12 Å. The MD trajectory snapshot was saved after every 1 ns; thus, 1000 frames for every simulation were collected for further analysis.

Distance between Domains

The center of mass of HNH, RuvC, REC1, REC2, and REC3 domains in both the Cas9 forms was calculated using the pdb_centermass script of pdbtools (https://github.com/harmslab/pdbtools). Additionally, the COM of Topo and CTD in SpCas9 and PI and WED domains in FnCas9 protein were also determined using the same script. To quantify and compare the domain movement due to gRNA binding, the distance between the center of mass of every domain pair during the simulation time (200-1000ns, stable part of trajectory) was calculated to characterize relative movements of intramolecular domains (due to gRNA binding) as a function of simulation time. Interdomain distances in both free and gRNA-bound forms for both the systems were averaged. The difference (Dist-Free form - Dist-bound form) in these averaged distances was determined for every domain pair. A positive delta indicates that the domains came closer, and vice-versa. Normalized delta distances were taken as input to form a clustermap to illustrate the relative movement of domains due to the binding of gRNA. A clustermap uses hierarchical agglomeration clustering to cluster similar columns together. The order of clustering forms a dendrogram. These values were standard scaled to 1 to make a comparison between both Cas9s.

Domain structural transitions and fluctuations

The propensity of a residue to attain a particular secondary structural element was computed for 1000 frames(Fig S4). A secondary structure was assigned to those residues which had a greater than 70% probability to exist either as a helix, turn, strand, or coil throughout the simulation. Those residues that did not conform to any particular secondary structure throughout our simulation (existed as a secondary structure for < 70% of times throughout simulation) were also quantified and referred to as dynamic residues. The percentage of these dynamical residues that make up each of the functional domains of the protein. To determine regions with high flexibility, the percentage of residues existing in coil conformation in more than 70% of the simulation time were calculated for every domain in Apo and gRNA-bound form for both the systems. Domain-wise atomic fluctuations were studied by plotting root-mean-square fluctuation for every domain. The

average RMSF of SpCas9 and FnCas9 proteins in free and bound form were plotted.

Principal component analysis

Principal component analysis (PCA) is a useful method for assessing trajectory data from MD simulations and identifying the essential dynamics Wan *et al.* (2013, 2015). Based on the calculation of the covariance matrix, this method applies a linear transformation to project the original high-dimensional representation of macromolecular dynamics into the low-dimensional space. The covariance matrix of the protein backbone atoms was built and diagonalized using gmx covar method of GROMACS 5.1 package Abraham *et al.* (2015) to investigate the dominant motions of free and gRNA bound forms of Cas9. Only a very small number of eigenvectors (modes of fluctuation) contribute significantly to the overall motion of the protein. This can be seen from the eigenvalues contained in eigenval.xvg. The first principle component (PC1) corresponds to the system's largest amplitude motion, and the system dynamics along the PC1 is generally called "Essential Dynamics". The most dominant mode, containing 100 conformations along eigenvector 1 was saved in PDB format and visualized in Normal Mode Wizard of VMD Humphrey *et al.* (1996).

MMGBSA Calculations

The Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method Genheden and Ryde (2015) is a popular technique for determining binding free energies between proteins and ligands. gmx_MMPBSA Valdés-Tresanco *et al.* (2021) was used to calculate the enthalpy contribution to binding free energy. The GB (OBC) model (igb=2) was used to calculate the solvation energy. A total of 700 snapshots, equally distributed in the last 50ns of the SpCas9+RNA and FnCas9+RNA trajectory were selected to calculate the enthalpy contribution using MM/GBSA. The single trajectory approach was used to calculate the binding free energy. Molecular mechanics potential energy (electrostatic and van der Waals energies), the free energy of solvation (polar and nonpolar solvation energies), and the energetic contribution of each residue and nucleotide to total binding energy were obtained for both SpCas9 and FnCas9 proteins with gRNA. The gas-phase energies calculated using the Molecular Mechanics force field, are a sum of van der Waals (VDWAALS) and electrostatic (EEL) energies, whereas the solvation energy includes a polar and a non-polar component. The polar component is calculated using the implicit Solvent Model, i.e. Generalized Born (GB), (EGB) and the non-Polar component depends on solvent accessible surface area (ESURF). Free energies of solvation are estimated by applying Poisson-Boltzmann (PB) calculations for the electrostatic contribution and a surface-area-dependent term for the non-electrostatic contribution to solvation.

Electrostatic Calculation of Cas9 domains

The electrostatic contribution made by HNH, RuvC, REC1, REC2, and REC3 domains were calculated using the Adaptive Poisson-Boltzmann Solver (APBS) Jurrus *et al.* (2018). Prior to using APBS, the standalone version of PDB2PQR Dolinsky *et al.* (2004) was used to assign charge and radius parameters for the CHARMM force field to all the domain structures. APBS uses iterative solvers to solve the nonlinear algebraic equations resulting from the discretized Poisson-Boltzmann equation with a fixed error tolerance of 10÷6. The electrostatic values for the centroid coordinates of each of these domains were then extracted from its corresponding electrostatic potential file using the multivalue script (provided in APBS).

Gibbs Free energy landscape

The free energy landscape (FEL) is used for describing conformation exchange in biomolecular processes such as molecular recognition, folding, and aggregation. The population and stability of functionally different states are determined by free energy basins and their depths, while the inter-basin barriers correspond to the transient states that connect them. gmx sham module of GROMACS was used to compute Gibbs Free Energy using the Boltzmann equation. The relative free energy between two states is given by;

$\Delta G = -KbT*\ln[P1(X)/P2(X)]$

where Kb is the Boltzmann constant, T is absolute temperature, X stands for the reaction coordinate and P(X) is the probability distribution of the system along the reaction coordinate. Based on the PCA calculation, PC1 and PC2 were chosen as reaction coordinates in this analysis. Frame-wise free energy profiles of all four Cas9s.

Computational Alanine Scanning

Computational Alanine Scanning is an energy-based method to calculate the effects of alanine mutations on the binding free energy of a protein-gRNA complex. This alanine substitution removes the side chain from the selected residue which aids in the assessment of the role of a side chain functional groups at a particular position. It is used to determine the significance of a residue in protein-gRNA binding. The principle followed is:

$\Delta\Delta G_{(CAS)} = \Delta Gbind_{(Ala)} - \Delta Gbind_{(Wild\text{-}type)}$

In this formula,$\Delta\Delta G_{(CAS)}$ is the binding free energy difference between $\Delta Gbind_{(Ala)}$ (residue substituted with Alanine) and $\Delta Gbind_{(Wild\text{-}type)}$. This analysis was performed to evaluate the potential of hotspot residues in gRNA binding. 50 frames from the last 10 ns part of the trajectory were considered as input for both the systems.

3.2 Results and Discussions

3.2.1 Conservation Analysis reveals the diversity of Sp- and FnCas9 proteins.

The results from ConSurf Ashkenazy *et al.* (2016) revealed that bridge helix (BH) and the HNH domains were most conserved (SpCas9: 44% and 41%; FnCas9: 43% and 35%) and the REC2 domain was found to be least conserved (SpCas9: 12%; FnCas9: 7%) in both the proteins shown in Figure 3.1A. In Figure 3.1B, the conservation scores (Color codes) assigned by ConSurf were used to illustrate residue-wise and domain-wise conservation. Since the bridge-helix is an arginine-rich motif that bridges the nuclease and REC lobe, hence, the percentage of charged residues in this region is higher in both the proteins (SpCas9 (BH-Charged residues): 22.8% and FnCas9 (BH-Charged residues): 25%). In all the other domains, non-polar residues have shown a major involvement in conservation. Pairwise-sequence alignment of SpCas9 and FnCas9 sequences was performed using EMBOSS, which implements Needleman-Wunsch Algorithm to find the optimum alignment (including gaps) of two sequences along their entire length Rice *et al.* (2000). 23.9% residues in SpCas9 were conserved with the FnCas9 sequence.

3.2.2 gRNA-binding to Cas9 proteins stabilizes the complex.

The atomic fluctuations and structural stability of the SpCas9 and FnCas9 systems during MD simulations were examined by monitoring the backbone root mean square standard deviation (RMSD) with respect to their starting structures as a function of simulation time. All the systems exhibited a stable deviation after 200 ns (200-1000 ns considered a stable part of the trajectory). Various geometrical properties, including the number of intra-protein hydrogen bonds, solvent accessible surface area (SASA), intermolecular hydrogen bonds between Cas9 protein and gRNA, and radius of gyration, were calculated from the equilibrium MD trajectories (Figure 3.1A,E). Figure 3.2C list the comparison between the average values of these geometrical properties in bound and unbound form for both the Cas9 systems. Time evolution of these properties was computed (Figure 3.1) revealing some subtle differences in geometrical properties in both cases. In the case of SpCas9, the average solvent accessibility (SASA) (Figure 3.1-C) and the average number of intra-protein hydrogen (Figure 3.1-D) bonds increased after gRNA-binding with a small increase in the compactness of the structure(Figure 3.1-B). The SASA plot (Figure 3.1-C) and average SASA values listed in Table 1 (3.2) suggest an increase in solvent accessibility due to gRNA-binding leading to exposure of hydrophobic residues and loss of compactness/ folding of the protein. In the case of FnCas9, a similar observation was made with an increased radius of gyration (Figure 3.1-F) and SASA after gRNA binding (Figure 3.1-G), while the number of intra-protein hydrogen bonds (Figure 3.1-H) decreased after gRNA-binding.

These observations suggest that gRNA-binding has stabilized the system (increased NHB); however, a reduced the compactness of the structure (increased ROG). This has led to increased

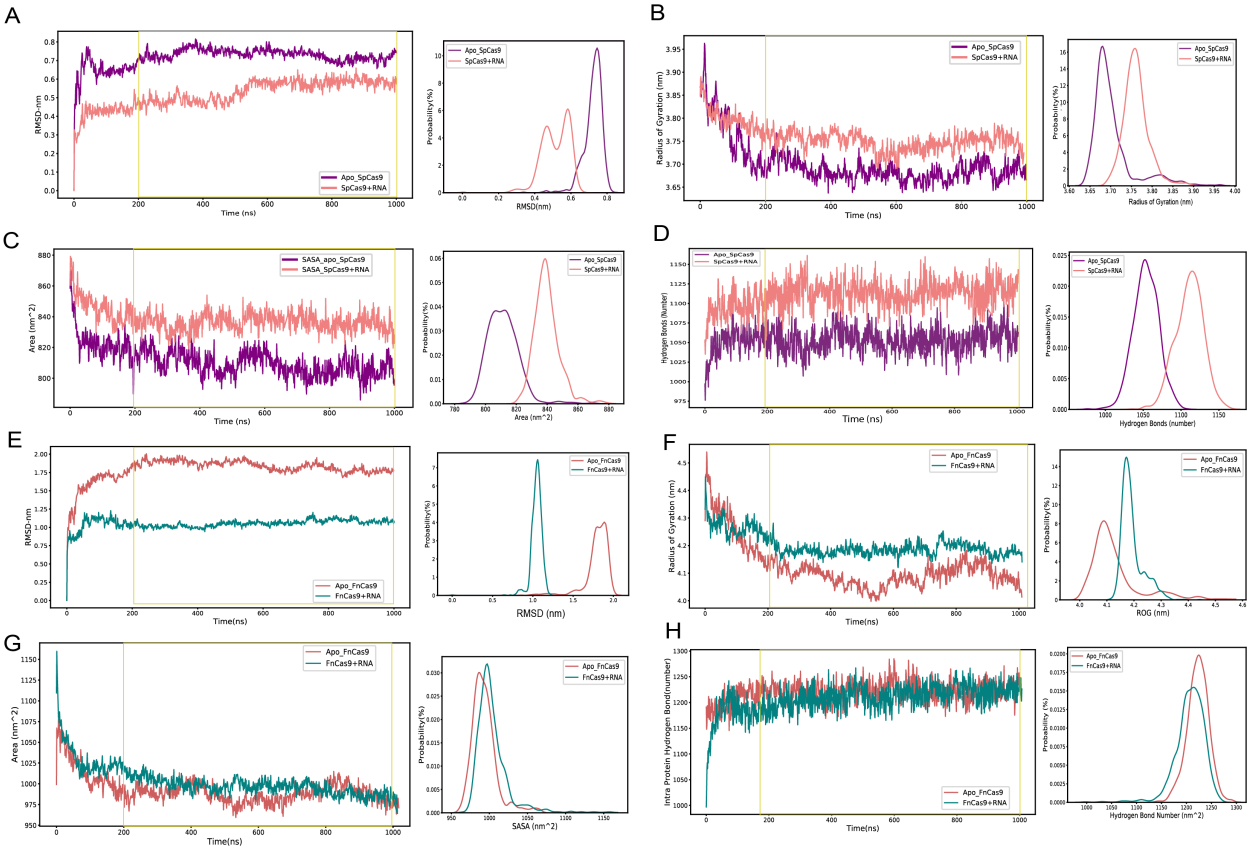
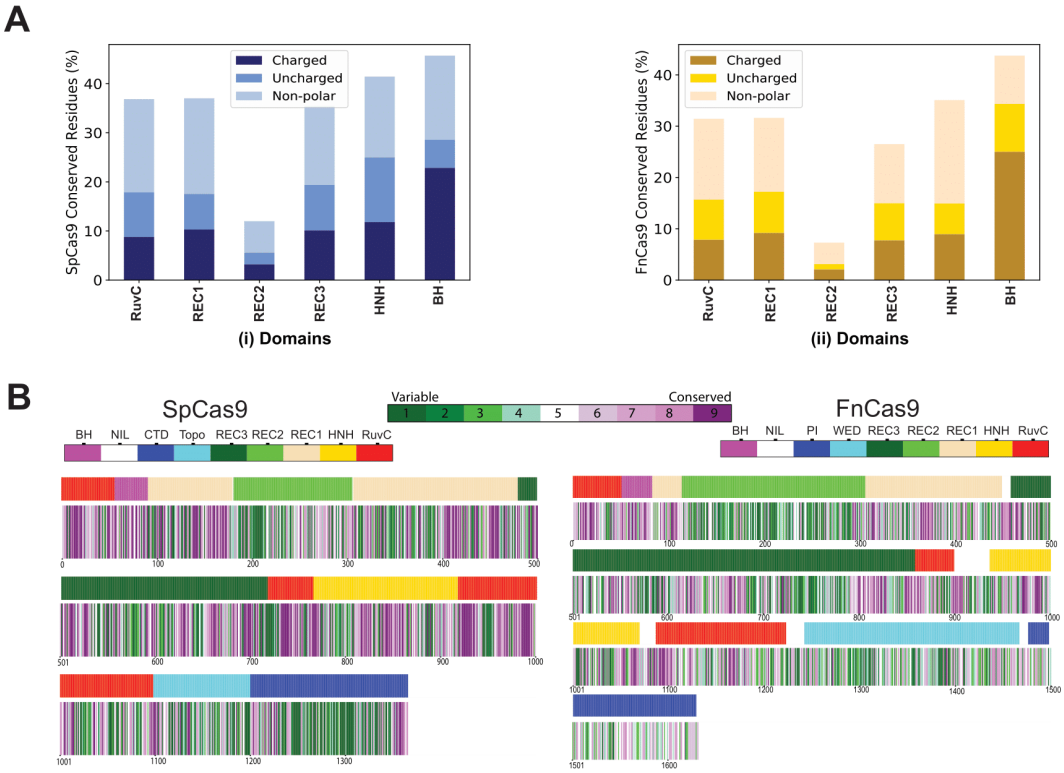


Figure 3.1: Figure: A. [Left] RMSD of backbone atoms over time for Apo-SpCas9 (violet) and RNA-bound SpCas9 (light-coral). [Right] RMSD probability distribution for Apo-SpCas9 and RNA-bound SpCas9. B. [Left] Radius of Gyration evolution for Apo and RNA-bound SpCas9. [Right] Probability distribution of Radius of Gyration for Apo-SpCas9 and RNA-bound SpCas9. C. [Left] Solvent accessible surface area (SASA) over time for Apo-SpCas9. [Right] SASA probability distribution for Apo and RNA-bound SpCas9. D. [Left] Time evolution of intra-protein hydrogen bonds for Apo-SpCas9 and RNA-bound SpCas9. [Right] Probability distribution of intra-protein hydrogen bonds for Apo and RNA-bound SpCas9. E. [Left] RMSD over time for Apo-SpCas9 (violet) and RNA-bound SpCas9 (light-coral). [Right] RMSD probability distribution for both. F. [Left] Radius of Gyration over time for Apo and RNA-bound SpCas9. [Right] Probability distribution of Radius of Gyration for both. G. [Left] SASA over time for Apo-SpCas9. [Right] SASA probability distribution for Apo and RNA-bound SpCas9. H. [Left] Intra-protein hydrogen bonds over time for Apo and RNA-bound SpCas9. [Right] Probability distribution of intra-protein hydrogen bonds for both.

solvent accessibility due to gRNA-binding and exposure to hydrophobic residues.

Residue-wise, RMSF is an often-used measure of conformational variance which highlights the regions of high mobility. To evaluate and compare the structural flexibility between free and gRNA bound forms of SpCas9 and FnCas9, per-residue root means square fluctuation (RMSF) was computed for the stable part of trajectory (200-1000 ns) and reported in Figure 3.3A for SpCas9 and Figure 3.4A for FnCas9. In Figure 3.3A, we observe that the RMSF curve for Apo-SpCas9 (violet) and SpCas9 with gRNA (light-coral) had similar atomic fluctuations with average RMSF, 0.26 nm and 0.24 nm. These RMSF values were mapped onto the protein structure to



C

Table 1: Geometrical Properties for stable part of SpCas9 trajectory (200-1000ns)

System	Average RMSD (nm)	Average RMSF (nm)	Average ROG (nm)	Average SASA(nm²)	Average-HBonds (count)	Intermolecular HBonds Spcas9-RNA
Apo-SpCas9	0.73	0.26	3.68	808.89	1053.22	90.613
SpCas9 + RNA	0.50	0.24	3.76	839.73	1108.82	

Table 2: Geometrical Properties for stable part of FnCas9 trajectory (200-1000ns)

System	Average RMSD (nm)	Average RMSF (nm)	Average ROG (nm)	Average SASA(nm²)	Average-HBonds (count	Intermolecular HBonds Spcas9-RNA
Apo-FnCas9	1.84	0.46	4.09	987.72	1224.58	131.12
FnCas9 + RNA	1.04	0.30	4.20	1007.67	1202.93	

Figure 3.2: A. Grouped Bar plot showing the percentage of conserved residues (residues with color codes $\geq$8) in SpCas9 (PDB ID: 4CMP) (on the left) and FnCas9 (PDB ID: 5B2Q) (on the right) domains and the contribution made by charged, uncharged and non-polar residues in conservation. B. Residue and domain-wise conservation of residues in SpCas9 (on the left) and FnCas9 (on the right). The color schemes followed for domains and conservation color codes are added to the color bar on the top. C. Average geometrical properties calculated from stable part of MD trajectories for SpCas9 and FnCas9 systems.

find regions of higher flexibility. Regions in REC1, Topoisomerase, and RuvC-III domains were found to have higher fluctuation (RMSF $>$ 0.6 nm) in unbound form (shown in Red color in Figure 3.3B), whereas only the Topoisomerase domain and some portions of CTD domains were found to be highly fluctuating in gRNA bound form of SpCas9, indicating lesser fluctuating regions post-gRNA-binding.

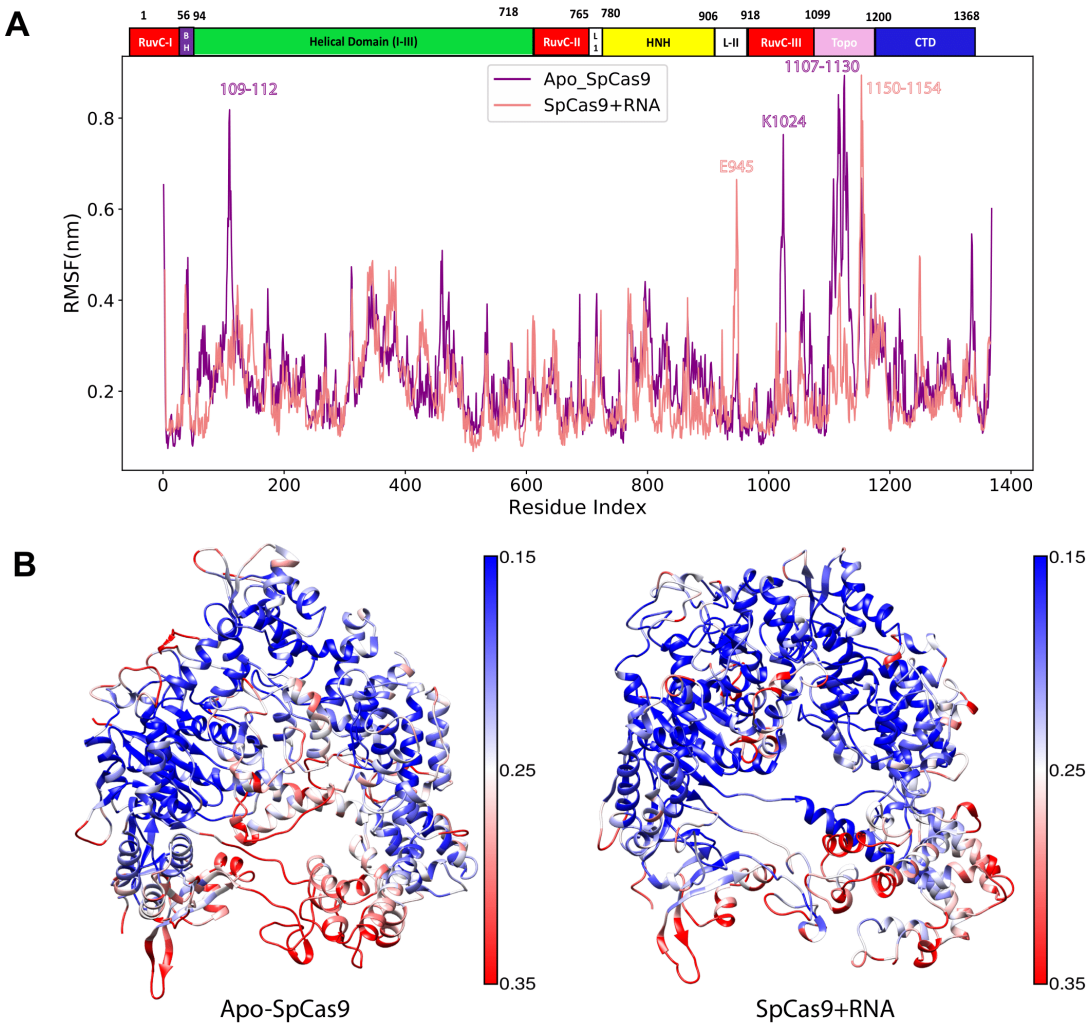


Figure 3.3: Comparison between the structural flexibility of Apo-SpCas9 and SpCas9 with gRNA. (A) Per-residue RMSF profiles were calculated from MD trajectories of Apo-SpCas9 (violet) and SpCas9+RNA (coral). The residues corresponding to every domain in the SpCas9 structure are shown. (B) The structural representations of Apo-SpCas9 and SpCas9+RNA mapped with per-residue RMSF values generated using UCSF Chimera. The color ranges from blue to red and denotes that RMSF varies from the lowest to the highest values.

The atomic fluctuations in Apo-FnCas9 (indian-red) were slightly higher than the gRNA-bound form (teal) with average RMSF, 0.46 nm and 0.30 nm (Table 2 in Figure 3.2C). The regions with high mobility (RMSF $>$ 0.6 nm) were determined by mapping RMSF values onto the FnCas9 protein structure. It was found that regions in the REC lobe, specifically segment of REC2-REC3 domains (residues 218-233, 727-791), HNH (residues 950-998; 1004-1065) were highly fluctuating in the free-FnCas9 form (shown in red color) and only some regions of REC2 domain (residues 216-224), HNH domain (residues 953-961; 1004-1065), and PI domains (residues 1565-1572) had high-mobility in the gRNA bound form of FnCas9 as shown in Figure 3.4. Both states of SpCas9 were equally flexible, whereas Apo-FnCas9 was relatively more flexible than its gRNA-bound form. On comparing, the radius of gyration, RMSD, and RMSF values, we can say that FnCas9 is more volatile than SpCas9 as it had a higher average radius of gyration and RMSF values.

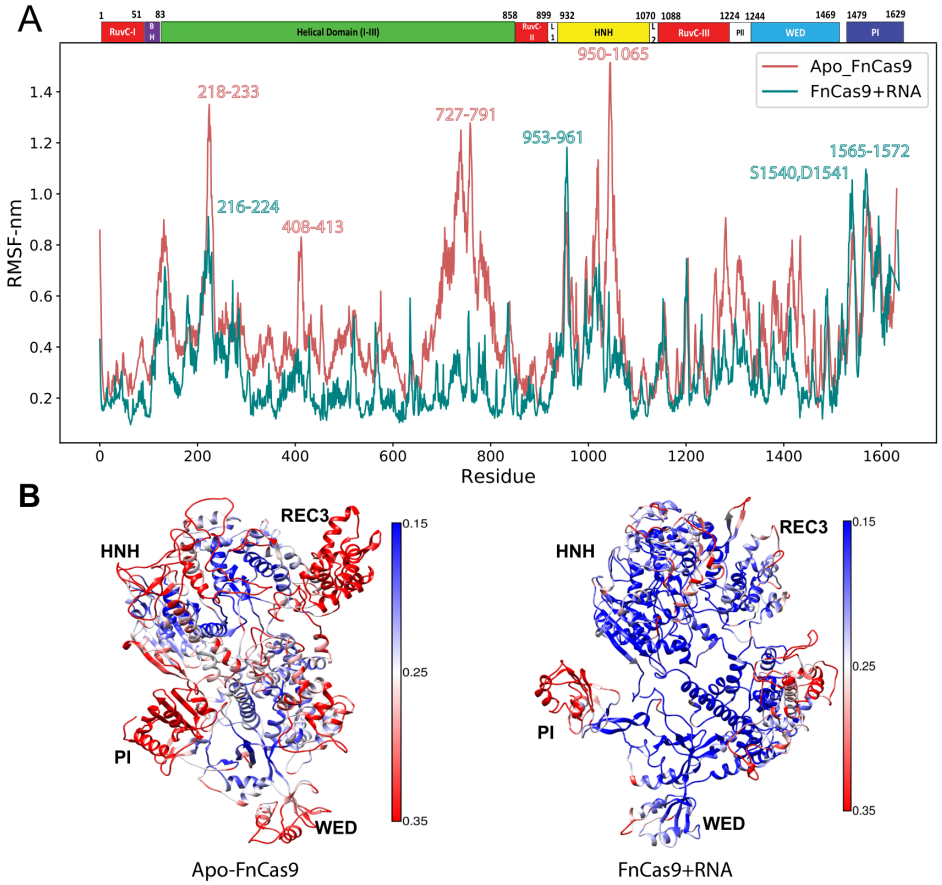


Figure 3.4: Comparison between the structural flexibility of Apo-FnCas9 and FnCas9 with gRNA. (A) Per-residue RMSF profiles were calculated from MD trajectories of Apo-FnCas9 (Indian-red) and FnCas9+RNA (teal). The residues corresponding to every domain in the FnCas9 structure are shown in the top horizontal bar. (B) The structural representations of Apo-FnCas9 and FnCas9+RNA mapped with per-residue RMSF values generated using UCSF Chimera. The color ranges from blue to red and denotes that RMSF varies from the lowest to the highest values.

3.2.3 Relative Domain Movement and Domain Structural Deformations are involved in the binding of gRNA to Cas9 proteins.

Many structural and biophysical studies have reported about the important conformational changes in Cas9 protein due to nucleic acid binding Jiang and Doudna (2017); Palermo *et al.* (2016*b*, 2017). Upon DNA binding, the HNH domain undergoes a significant conformational change leading to activation of Cas9 protein for catalysis of TS Sternberg *et al.* (2015); Palermo (2019*b*). Enhanced MD simulations of pre-activated CRISPR–Cas9 (5F9R.pdb) ($\sim$ 16 μs) revealed that the tight-coupling between REC lobe and HNH domain aided in bringing the Cas9 to the activated state. The conformational transitions observed in the REC lobe disclosed how the recognition domains 'sense' nucleic acids, 'regulate' the HNH conformational change, and ultimately 'lock' the HNH domain at the cleavage site, contributing to its catalytic activation Palermo *et al.* (2018). These studies justify the significance of conformational dynamics of CRISPR–Cas9 in DNA catalysis, thereby improving its genome editing abilities.

The relative movements of intramolecular domains due to the binding of gRNA were determined by calculating distances between the center of mass of each domain as a function of simulation time for both the Cas9 forms.

The center of mass of HNH, RuvC, REC1, REC2, REC3, PI, and CTD domains were computed for both the systems (SpCas9 and FnCas9), and the distances between COM (centre of mass) of every domain pair for the simulation length were calculated. The domain movement due to gRNA binding was quantified by plotting a clustermap using the difference in average domain distance (200-1000 ns) in free and gRNA bound form. Upon gRNA binding in SpCas9, we observed that the domain movement by HNH-REC2 domains, CTD-Topo domains, and REC3-RuvC domains was similar (Figure 3.5A). The HNH and REC2 domains moved closer to REC3- RuvC domains and moved away from the REC1 domain after gRNA binding in SpCas9 form. Similarly, CTD and Topo moved closer to REC2, and REC3, and moved away from RuvC and REC1 domains. The RuvC and REC3 domains moved closer to REC2 and HNH domains and moved farther from Topo domains. In Fncas9, the PI, WED, and RuvC domains moved closer to each other, and REC3 and HNH domains, and moved away from the REC2 domain (Figure 3.5B). REC1 and REC2 domains have shown a similar domain movement by moving away from RuvC, PI, and WED domains. The FnCas9 protein has shown larger domain deviations as compared to SpCas9, as the domain deviation scale was larger in FnCas9.

The structural deformation of each of these domains in terms of domain flexibility and domain-wise structural transitions was studied by computing the domain-wise RMSF (Figure 3.5C and Figure 3.5D). The atomic fluctuations of every domain in the free form of SpCas9 were observed to be greater than in the gRNA-bound form. This observation was in agreement with other observations (in Figure 3.2 and Figure 3.3), suggesting increased stability in the presence of nucleic acids. The Topo and REC1 domains were found to be highly mobile in both the forms of SpCas9. The domain-wise average RMSF was calculated for Apo-FnCas9 and FnCas9 with gRNA for the stable part of the trajectory and plotted in Figure 3.5B. A similar observation was made in the case of FnCas9 too, where the atomic fluctuations of domains in the free form of FnCas9 were greater than the gRNA bound form for RuvC, REC1, and REC2 domains, whereas the average RMSF of the HNH domain was higher in gRNA bound form than in the free form. The PI and REC3 domains were highly mobile in the pre-gRNA-bound form of FnCas9.

To evaluate the conformational flexibility of these domains, the percentage of residues existing in coil conformation for more than 70% of the simulation time was calculated for every domain in Apo and gRNA-bound form for both the systems in SpCas9 form (Figure 3.5C) and in FnCas9 form (Figure 3.5D). In SpCas9, the percentage of residues existing as a coil for more than 70% times, decreased after gRNA binding for all the domains except in the HNH and the REC2 domains. In FnCas9, the percentage of coils increased after gRNA-binding in RuvC, REC1, and PI domains.

The propensity of a residue to exist as a helix, sheet, turn, or coil was computed for the stable part of the trajectory. Also, dynamic residues with high order of transition between secondary

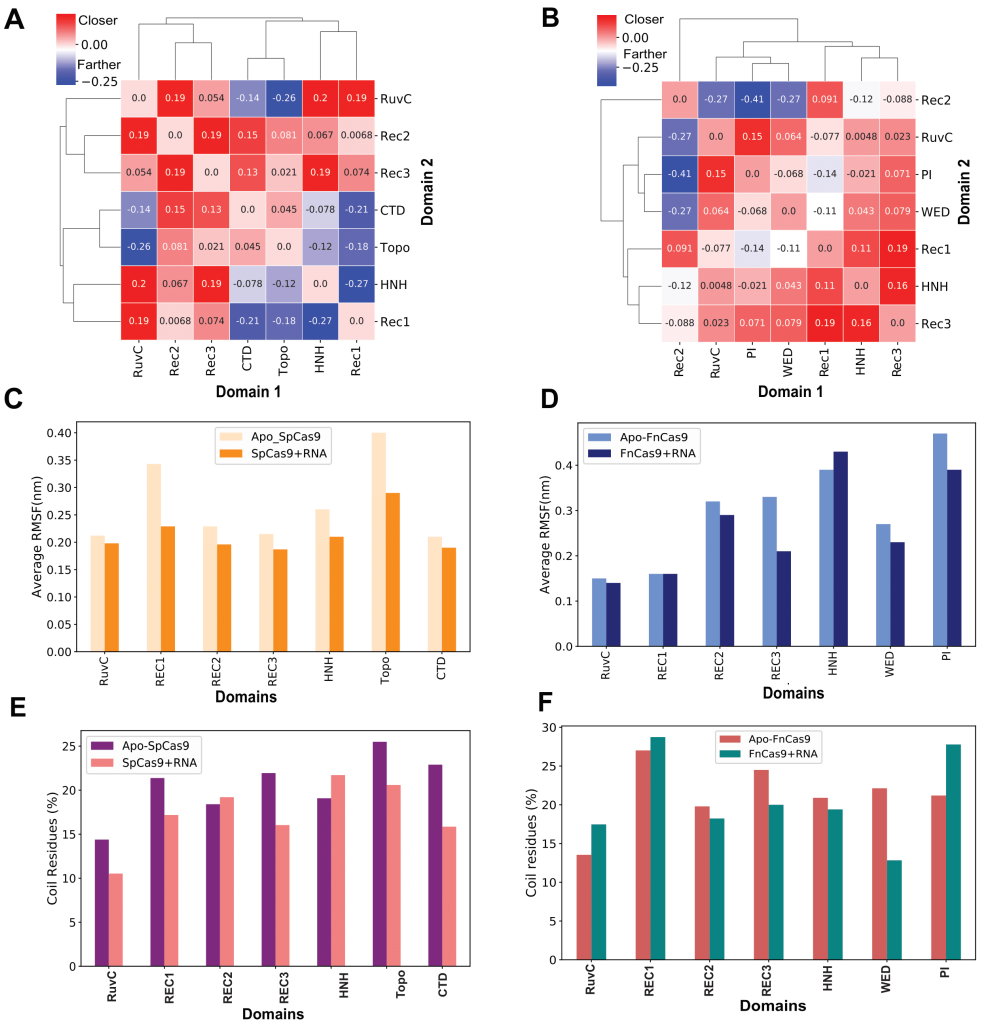


Figure 3.5: Domain Structural movements and transitions in Clustermap shows domain movement between domains on the X and Y-axis. Red boxes suggest domains moved closer and blue boxes are for domains that moved away. The color intensity depends on the magnitude of domain movement. (A) Clustermap for SpCas9 (B) Clustermap for FnCas9 Grouped bar plot with average RMSF values (200-1000ns) corresponding to RuvC, REC1, REC2, REC3, and HNH domains in the free and bound form of (C) SpCas9 system and (D) FnCas9 system.

structures were annotated (See Materials and Methods for details). The secondary structure assignment of residues in Apo-SpCas9 and SpCas9 with gRNA are shown in Figure 3.6-A(i) and Figure 3.6-B(i), and Apo-FnCas9 and FnCas9 bound to gRNA are shown in Figure 3.6-C(i) and Figure 3.6-D(i), respectively. The probability of residue to exist as a helix, sheet, or coil for 1000 frames was calculated and plotted to see regions of structural deformations at the residue level (SpCas9-Figure 3.6-A(ii) and Figure 3.6-B(ii); FnCas9-Figure 3.6-C(ii) and Figure 3.6-D(ii)).

The secondary structure transition of residues among several conformations of proteins contributes to the structural flexibility of the proteins. The percentage of these dynamic residues that make up each of the functional domains of both the forms of SpCas9 and FnCas9 proteins are shown in Figure 3.7A and Figure 3.7B. The percentage of dynamic residues has increased after

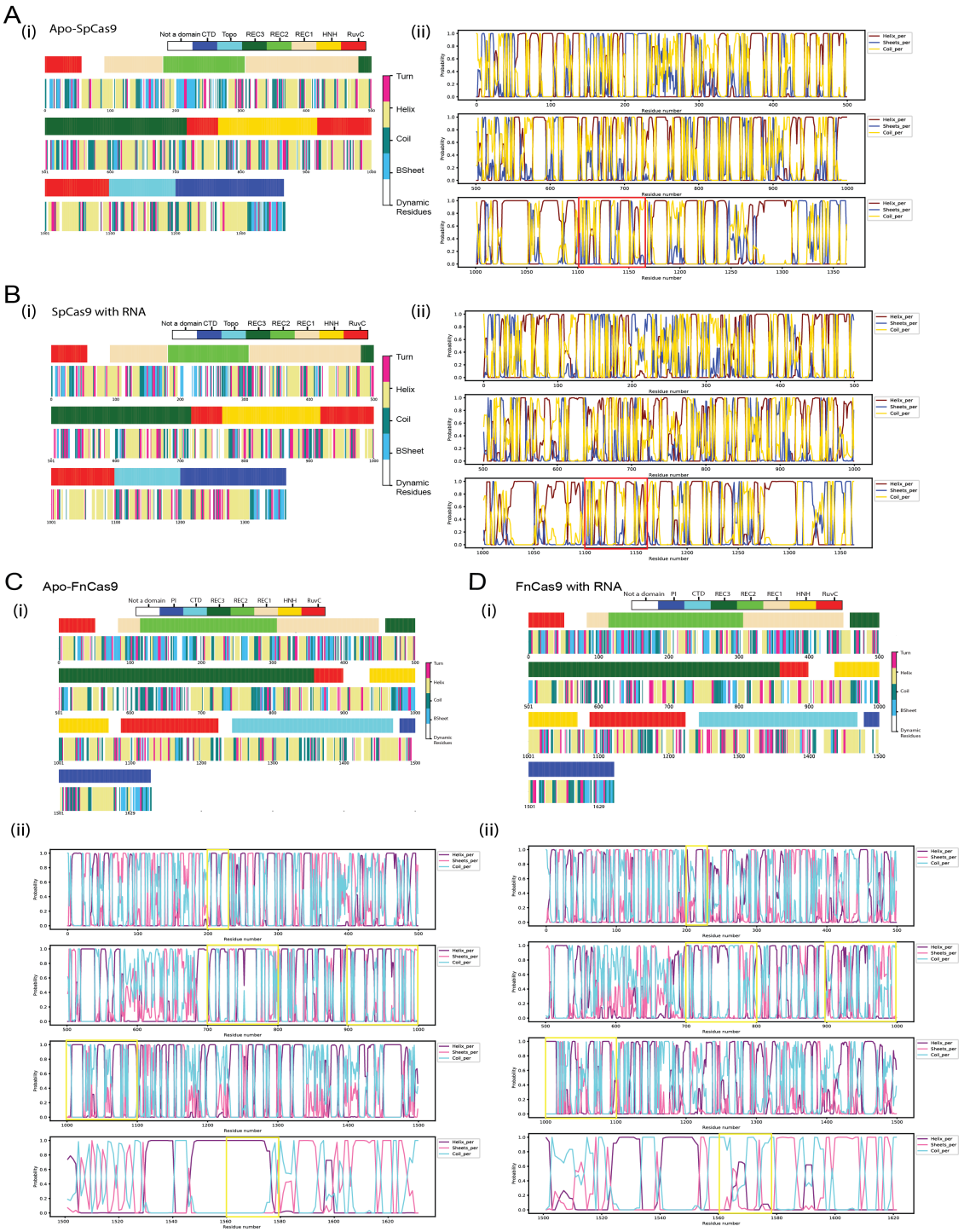


Figure 3.6: Secondary structure and propensity analysis of Apo- and RNA-bound SpCas9 and FnCas9. A (i) and B (i) show residue-wise and domain-wise secondary structure assignment (present in >70% of frames in stable parts) for Apo-SpCas9 and SpCas9-RNA, respectively. A (ii) and B (ii) display secondary structure propensity per residue across 1000 frames for Apo-SpCas9 and SpCas9-RNA, with Helix (Brown), Sheet (Blue), and Coil (Yellow) probabilities. High fluctuation regions (RMSF) are boxed in red. C (i) and D (i) show the residue-wise and domain-wise assignment for Apo-FnCas9 and FnCas9-RNA. C (ii) and D (ii) display secondary structure propensity per residue across 1000 frames for Apo-FnCas9 and FnCas9-RNA with Helix (Violet), Sheet (Pink), and Coil (Cyan) probabilities.

gRNA binding for all the domains in both Cas9s. In the case of Spcas9, residues with higher structural transitions were found to predominantly exist in the REC lobe while the HNH domain had the least dynamic residues. The REC2 domain had shown a drastic increase in dynamic residues from 15% to 35% after gRNA binding, and on the contrary, HNH and CTD domains had shown less percentage of dynamic residues. This observation was in line with conservation analysis where the HNH domain was the most conserved, and the REC2 domain was least conserved. The conformational flexibility of a protein region also depends on the proportion of regions with NMR chemical shifts of a random-coil; regions that lack significantly ordered secondary structure (as determined by CD or FTIR) Uversky *et al.* (2000). A similar observation was made in the case of FnCas9, the percentage of dynamic residues increased after gRNA binding. Here, REC lobe contained the highest percentage of dynamic residues after gRNA binding. After applying 70% cut-off, it was observed that the protein structure was predominantly constituted by helices (30-40% residues), followed by the dynamic residues (16-23% residues) and coils (16-20%), while sheets, turns, and bends were the least favored (4-12% respectively). Interestingly, RMSD and geometrical properties of the system during the simulation revealed that the gRNA binding was stabilizing the structure, however, the secondary structure analysis revealed a decrease in percentage helicity (Apo-SpCas9: 42.03%, SpCas9+RNA: 37.21%; Apo-FnCas9: 38.73%, FnCas9+RNA: 30.95%) and increase in the percentage of dynamic residues (Apo-SpCas9: 16.74%, SpCas9+RNA: 22.15%; Apo-FnCas9: 17.03%, FnCas9+RNA: 21.39%) due to the binding of gRNA. The Topo and REC1 domains had higher residues as coils which could be a reason behind the high average RMSF/fluctuations.

3.2.4 Larger domain deviations are observed in SpCas9 protein than in FnCas9.

The principal component analysis was performed to capture the essential motions of the simulated systems. To investigate the dominant motions of free and gRNA-bound forms of Cas9 along the first principle component (PC1), corresponding to the system's largest amplitude motion, 100 conformations of the system along eigenvector1 were stored in PDB format and visualized in Normal Mode Wizard of VMD Humphrey *et al.* (1996). The eigenvalues indicated fluctuations of the eigenvector in the hyperspace. The overlap between principal components and the coordinates of the trajectory was determined for free and bound states of SpCas9 and FnCas9. Figure 3.7-C and Figure 3.7-D show the conformational sampling of free and bound forms of SpCas9 and FnCas9, respectively, in the essential subspace by projecting the backbone atom, showing the tertiary conformations along eigenvector 1 and 2. The 2D projection clearly depicted that free forms of Cas9 cover a wide range of phase spaces as compared to gRNA-bound forms. The decrease in the overall motion of Cas9 forms due to gRNA binding indicates the stability of the structure in the presence of nucleic acids.

In Figure 3.7E, the dynamics of Apo-SpCas9 protein along the first principal mode of motion (PC1) are shown. REC1, HNH, and CTD domains have shown larger domain movements. The

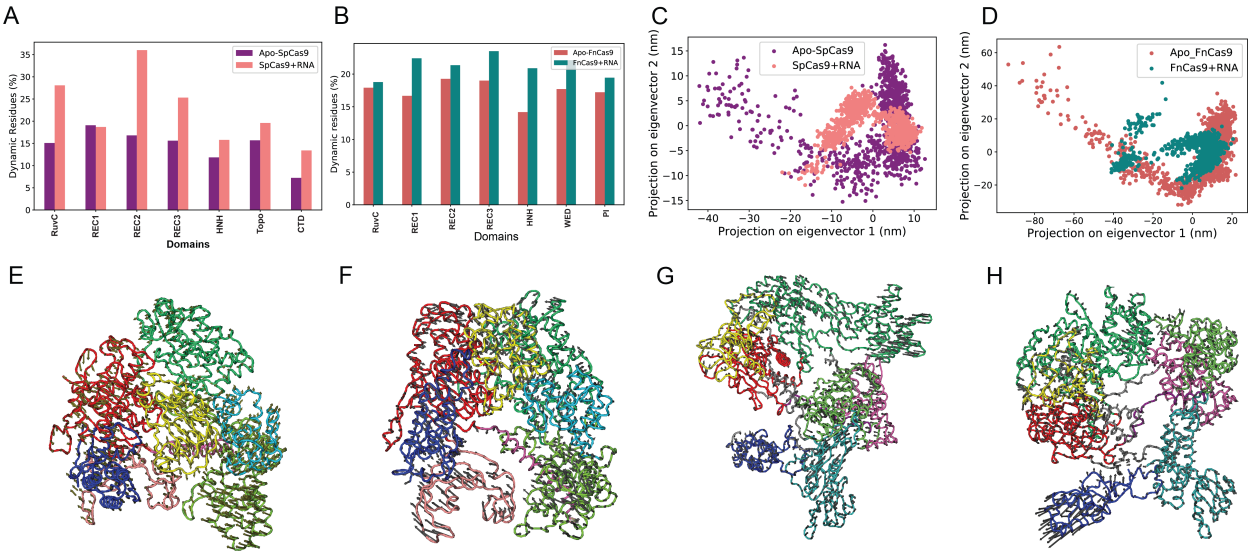


Figure 3.7: Domain Structural movements and transitions in FnCas9. Domain-wise percentage of dynamic residues ($>$ 70% part of trajectory) in (A) Apo-SpCas9 and SpCas9 with gRNA and in (B) Apo-FnCas9 and FnCas9 with gRNA. 2D projection of backbone atoms in essential subspace along first two eigenvectors of (C) Apo-SpCas9(violet) and SpCas9 with RNA (lightcoral) (D) Apo-FnCas9 (indian-red) and FnCas9 with RNA (teal). Most dominant mode, containing 100 conformations along eigenvector 1 for (E) Apo-SpCas9 and (F) SpCas9 with gRNA (G) Apo-FnCas9 (H) FnCas9 with gRNA form and visualized in NMWiz of VMD. The arrows indicate the direction and relative amplitude of motions.

RuvC domain moved closer to the REC lobe. The HNH and CTD domains have moved towards the REC1 domain, and moved away from REC3 and REC2 domains. The dominant motions in the case of the gRNA-bound form of SpCas9 in Figure 3.7F, included REC1, CTD, and Topo domain movements. HNH and RuvC domains moved closer to the REC1 domain. The domain movements in the case of the free form of SpCas9 were larger than the gRNA-bound form.

The most dominant mode of motion of the Apo-FnCas9 form is shown in Figure 3.7G. The HNH, RuvC, WED and REC3 domains moved closer to the REC2 domain. The WED domain moved away from HNH and PI domains. In gRNA bound form (Figure 3.7H), major domain movement is shown by PAM interacting domain (PI domain), it moved away from the REC lobe. The reorganization of the PI domain after gRNA binding is indicative of the structural preparation of FnCas9 for the next step of gene editing, which requires scanning of PAM motif and binding with the DNA strands. In both the Cas9 forms, the large amplitude motions were observed in the Cas9 domains which are directly involved in the process of recognition and binding of nucleic acids, i.e. REC lobe, regions of HNH, PI, and CTD domains. The ribonucleic acid-binding has stabilized the complex. In the first mode of motion, SpCas9 protein has shown larger domain deviations as compared to FnCas9.

A three-dimensional free energy landscape (FEL) was built for the free and gRNA-bound form of both Cas9 systems using the projections of first (PC1) and second (PC2) eigenvectors to show protein stability in terms of Gibbs Free energy. Figure 3.8 shows the FEL values for Apo and

gRNA bound form of SpCas9 and FnCas9. In the case of Apo-SpCas9, there was only one main global free energy minimum region, indicating only one stable conformational state. The lowest energy conformer of Apo-SpCas9 was identified at 636 ns with local free-energy 0 kJ/mol (State A). The second low energy state in the same energy basin had the energy of 0.74 kJ/mol (State B). The inter-conversion from State A to State B had an energy barrier of 2.2 kJ/mol while from State B to State A had an energy barrier of 2.96 kJ/mol. Whereas the gRNA-bound form of SpCas9 had shown two low-energy basins with energy minima at 508 ns and 850 ns (G= 0 kJ/mol). The inter-conversion between these states had an energy barrier of 3.8 kJ/mol. The state transition barrier is slightly higher in the gRNA-bound form than in the free-form of SpCas9, suggesting that state transition is easier in Apo form than in gRNA bound form of SpCas9. Likewise, in Apo-FnCas9 and FnCas9 bound to gRNA, two energy basins were observed. The lowest energy conformation was extracted at 478 ns and 249 ns, respectively. In Apo-FnCas9, the lowest energy state was at 0 kJ/mol (State A), and the second low energy state had 1.6 kJ/mol (State B) of energy. The transition from State A to State B of Apo-FnCas9 had an energy barrier of 5.9 kJ/mol and from State B to State A, the energy barrier was 4.3 kJ/mol. In the gRNA-bound form of FnCas9, the lowest energy state had 0 kJ/mol energy, and other low energy states had 0.79 kJ/mol of energy. Their inter-conversion had an energy barrier of 3.96 kJ/mol (State A to State B) and 3.17 kJ/mol (State B to State A). Here, energy barriers for state transitions was lower in the gRNA bound of FnCas9 form than in its free form (Figure 3.8), suggesting switch between low energy states is easier in gRNA bound form than in the Apo form of FnCas9.

3.2.5 Charge potentials reveal key insights into the greater specificity of FnCas9 protein.

MM-PBSA/GBSA (Molecular mechanics/Poisson–Boltzmann (Generalized-Born)) is an end-state post-processing method to estimate binding free energies. This method has been proven to balance accuracy and computational efficiency, especially when dealing with large systems. In order to calculate the binding affinity of gRNA with SpCas9 and FnCas9 forms and to understand the role of electrostatics and the atomic level interactions of gRNA to the Cas9 protein, MM-GBSA calculations were performed over an ensemble of 700 snapshots extracted from 50 ns (stable short region of the MD trajectory (950-1000ns) using gmx_MMPBSA. Per-residue energy decomposition of MM-GBSA was used to retrieve the individual amino acid contribution to the binding energy of gRNA with Cas9 protein . The average of each energy component that contributes to the binding free energy of gRNA with SpCas9 and FnCas9 is shown in Table 3.9A. The guide RNA (SpCas9: 85-mer, FnCas9: 94-mer) had a lesser binding affinity with SpCas9 form (-655.4 kcal/mol) than with FnCas9 form (-1254.1 kcal/mol). In both cases, the gas phase energy, especially the electrostatic energy, had contributed predominantly to the binding energy of gRNA and protein. In the case of SpCas9 (Table 3.9A(i)), the contribution of electrostatic energy (-2377.7 kcal/mol) was less than the polar solvation energy (2452.0 kcal/mol). On the contrary, in FnCas9 system (Table 3.9A(ii)), the electrostatic interaction (-4880.2 kcal/mol) had a greater contribution

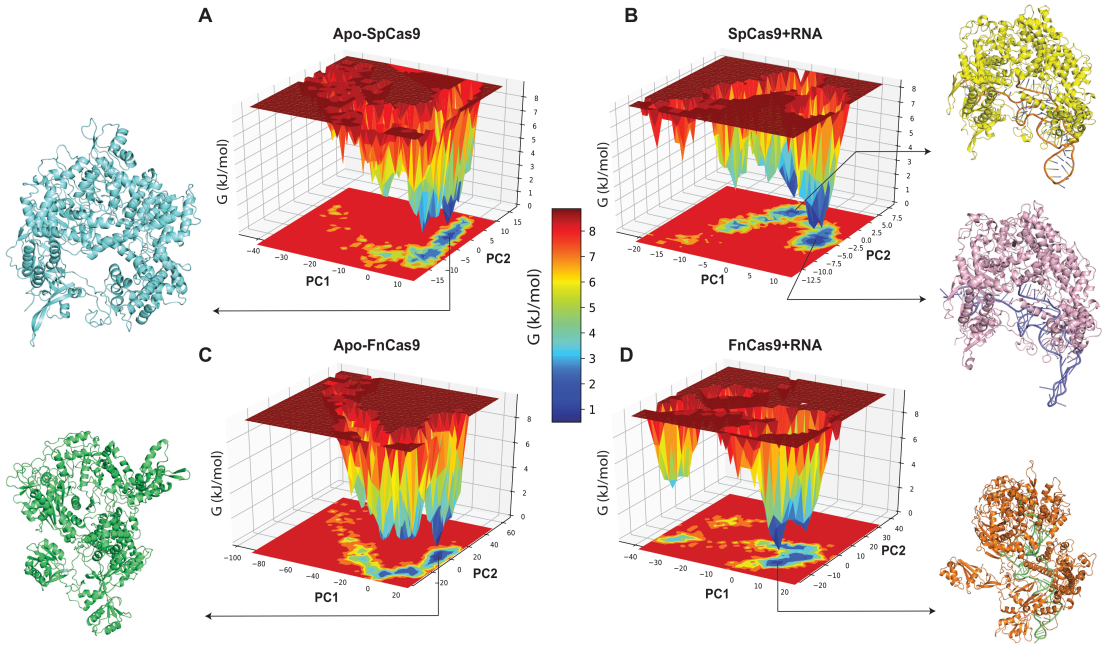


Figure 3.8: A. Free energy landscape generated using principal components (PC1 and PC2) for Apo-SpCas9 and SpCas9 with gRNA during 1000ns simulations.(B) Free energy landscape generated using principal components (PC1 and PC2) for Apo-FnCas9 and FnCas9 with gRNA during 1000ns simulations. Energy distribution is shown by the coloring pattern: Blue defines the conformational space with minimum energy (stable state) while red defines a conformational space with maximum energy (unstable state). Principal components are displayed as a contour map at the bottom of each FEL plot with similar color pattern like the energy landscape.

than the polar solvation energy (4748.2 kcal/mol). The energetic contribution of each residue was explored to identify the hotspot residues involved in the binding of Cas9 protein with gRNA. This analysis revealed that residues from the arginine-helix and REC1 domain play a pivotal role in the binding of gRNA with SpCas9 protein; however, only a few residues had shown a greater contribution to the binding energy (less than -6 kcal/mol). Many residues were found to be involved in the binding of gRNA (171) with FnCas9 protein; only those residues which have contributed more than -2 kcal/mol (< -2 kcal/mol) were considered important. These residues belonged to WED, REC3, REC1 domains, and the bridge-helix region of the protein. The energy decomposition distribution for both Cas9 forms was compared in Figure 3.9B. The boxplot of residue-wise energy decomposition revealed that both the Cas9 forms had nearly equal median energy values with a slightly greater inter-quartile energy range for FnCas9. Interestingly, there was one outlier, Phe105 (-8.3 kcal/mol) in SpCas9 and five outliers in FnCas9, namely, Arg401 (-7.6 kcal/mol), Arg63 (-7.8 kcal/mol), Arg62 (-8.6 kcal/mol), Arg55 (-10.7 kcal/mol), and Arg455 (-12.8 kcal/mol) with energy contribution greater than -7.5 kcal/mol. These residues were highlighted on the surface of SpCas9 and FnCas9 structures in sticks representation in Figure 3.9D. In FnCas9, all the outlier residues (energy < -7.5 kcal/mol) were arginines, three of which were a part of the bridge helix suggesting the pivotal role of arginine helix in gRNA recognition and binding.

Due to the strong negative electrostatic field associated with the gRNA phosphate backbone,

charged interactions are expected to be critical for correct protein-RNA binding and functioning. The MMGBSA results emphasized the significance of positively charged arginine residues located in close proximity to the gRNA-binding site. This observation can be correlated with the role of electrostatic interactions in gRNA binding. To elucidate the contribution made by each Cas9 domain in maintaining a suitable environment for gRNA binding, the electrostatic potential of each domain was evaluated using the APBS Baker *et al.* (2001) tool and compared with their gRNA-free and gRNA-bound forms in Figure 3.9C. A highly-positive electrostatic potential was shown by the HNH domain, and a highly-negative electrostatic potential was shown by the REC2 domain in the case of Apo-SpCas9. A drastic change in electrostatic potential was observed in the case of the REC2 domain of SpCas9 after gRNA binding, where the negative potential transformed to a positive potential. In the case of FnCas9, RuvC domain transformed its potential from electropositive to electronegative. On comparing the electrostatics of domain centers after gRNA binding, it was found that FnCas9 domains had high electrostatic potential as compared to SpCas9 (Positive potential in HNH, REC1, and REC3 domain and Negative potential in RuvC and REC2 domains), supporting the observation made in work by Acharya et al.Acharya *et al.* (2019*a*).

Computational Alanine Scanning is an effective method to determine the significance of side-chain functional groups of hotspot residues in protein-gRNA binding by substituting it to Alanine and then evaluating the change in binding affinity between Cas9 protein and gRNA in both the systems. The binding free-energy change due to Alanine mutation of hotspot residues (Spcas9 System: Phe105; FnCas9 System: Arg401, Arg63, Arg62, Arg55, and Arg455) was performed to assess their significance in gRNA binding. The $\Delta\Delta G_{(CAS)}$ for these residues is shown in Table 3.1. The alanine conversion of Phe105 decreased the binding affinity of gRNA with SpCas9 by 9.958 kcal/mol, and Arg455 decreased the binding affinity of gRNA with FnCas9 by 17 kcal/mol (similar values were observed for other hotspot residues), indicating an unfavorable alanine-substitution demonstrating the significance of these residues in gRNA binding.

Table 3.1: Computational alanine scanning result for hotspot residues involved in binding of gRNA with SpCas9 and FnCas9 protein. All values are in kcal/mol.

HotSpot Residues	$\Delta\Delta G_{(CAS)}$
Phe105 (SpCas9)	9.95 +/- 1.2
Arg55 (FnCas9)	15.23 +/- 1.5
Arg62 (FnCas9)	13.06 +/- 0.8
Arg63 (FnCas9)	13.15 +/- 1.8
Arg401 (FnCas9)	12.87 +/- 1.7
Arg455 (FnCas9)	17.00 +/- 1.7

Conclusions

Comparative analyses of gRNA-free and gRNA-bound forms of SpCas9 and FnCas9 suggested that the binding of gRNA has brought structural stability to the system. gRNA-binding has led to a decrease in structural compactness, indicating the opening of the structure and exposure of

A

Binding Energy (Average kcal/mol)

Model System	VDW	EEL	EGB	ESURF	ΔG gas	ΔG solv	ΔG Total
SpCas9 - gRNA	-649.69 +/- 61.43	-2377.69 +/- 23.96	2451.96 +/- 24.79	-79.98 +/- 1.45	-3027.39 +/- 65.44	2371.98 +/- 24.53	-651.40 +/- 61.10
FnCas9 - gRNA	-991.47 +/- 16.82	-4880.24 +/- 34.78	4748.24 +/- 33.60	-130.64 +/- 1.85	-5871.67 +/- 38.29	4617.59 +/- 33.84	-1254.07 +/- 17.07

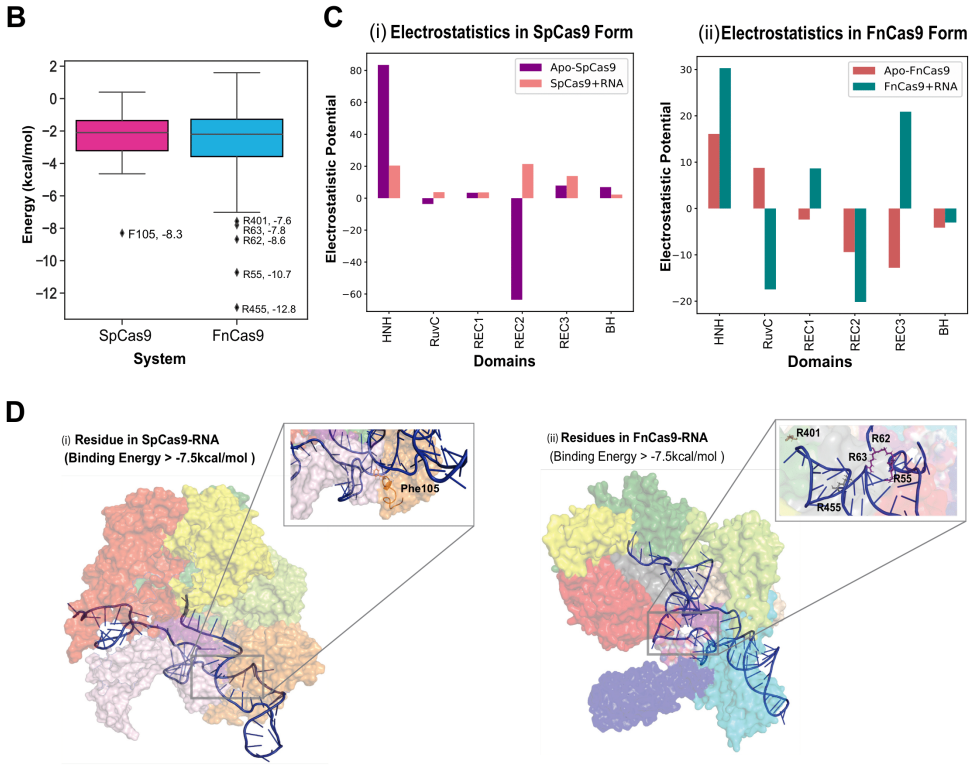


Figure 3.9: A. Bar plots indicating the free energy components (vdW, electrostatic, polar surface, SASA, obtained from MM-PBSA calculations for gRNA binding with SpCas9 (left) and FnCas9 (right). B. Box plot showing the distribution of per-residue energy contribution in the binding of gRNA with SpCas9 and FnCas9. C. Electrostatic potential of HNH, RuvC, REC domains, and bridge-helix in SpCas9 and FnCas9. D. Hotspot residues involved in gRNA binding shown in sticks representation in SpCas9 and FnCas9.

hydrophobic residues (3.2). The RMSF-profile for both states of SpCas9 had similar flexibility. Interestingly, in FnCas9, the Apo-form was more flexible than the gRNA-bound form (3.4). Residues in the REC2-REC3 domains (residues 218-233, 727-791) and the HNH (residues 950-998; 1004-1065) domain were highly fluctuating in the gRNA-free form, whereas the PI domain (residues 1565-1572) had high-mobility in the gRNA bound form of FnCas9. The FnCas9 was more volatile than the SpCas9 form, shown by the high radius of gyration, RMSD, and RMSF values. Domains with higher fluctuations had shown larger domain movement, suggesting a correlation between them (especially in Apo-FnCas9). As a consequence of gRNA-binding, a concerted series of domain movements occurred in SpCas9 (HNH-REC2, CTD-Topo, and REC3-RuvC) and

FnCas9 (HNH-REC3-REC1 and PI-WED-RuvC). The FnCas9 form had shown larger domain deviations as compared to SpCas9 as their difference in average domain distance range was higher than the latter. The increase in the percentage of dynamic residues observed in both Cas9s post-gRNA-binding implies the occurrence of higher structural deformations in those regions (Figure 3.6). In SpCas9, the REC2 domain contained a higher percentage of dynamic residues, and HNH had low dynamic residue percentage. This observation was in line with conservation analysis, where the HNH domain was most conserved, and the REC2 domain was least conserved. The high fluctuations and movement shown by the PI domain in the gRNA-bound form of FnCas9 could be due to higher structural transitions and increased coil percentage after gRNA-binding. In the free-energy landscape, clusters with frames of the initial part of the trajectory were structurally and energetically distinct. Energetically more stable states were observed in the gRNA-bound form in both the Cas9s. The binding free energy analysis revealed that gRNA is more tightly bound in FnCas9 form than in SpCas9. In SpCas9, the polar-solvation energy (EGB) had contributed predominantly to the binding energy, whereas in FnCas9 electrostatic energy (EEL) was the major contributor, disclosing the significance of electrostatic interactions in gRNA-FnCas9 binding. The per-residue energy decomposition in SpCas9 form revealed the significance of arginine-helix and REC1 domain in gRNA binding. Many residues have shown contribution towards the binding of gRNA in FnCas9, mainly belonging to WED, REC3, REC1 domains, and the bridge-helix region. Hotspot-residues (energy < -7.5 kcal/mol) with the greatest contribution towards gRNA binding were Phe105 in SpCas9 and Arg455, Arg401, Arg55, Arg62, and Arg63 in FnCas9. All hotspot residues were arginines in the case of FnCas9, three of which were a part of the bridge helix (R55, R62, R63), suggesting its pivotal role in gRNA recognition and binding.

This study talks about the sequential differences as well as changes in dynamics at the atomic level between Cas9 orthologs, SpCas9 and FnCas9. It sheds light on the role of bridge-helix residues and electrostatic interactions in FnCas9-gRNA binding. Cas9 domains play an essential role in bringing the protein into the catalytic-competent stage and, ultimately, in the DNA cleavage process. The structural changes and differences in domain rearrangement in both Cas9 orthologs show that the cleavage mechanism differs in these systems, which might explain why Cas9s have different specificity and sensitivity. To gain a better understanding about the cleavage mechanism at atomic level DNA bound ternary complex of SpCas9 and FnCas9 should be simulated and compared. Future research should focus on comparing the domain rearrangement and change in binding free energies in the presence of DNA and implementing high-level quantum-mechanical simulations to describe the formation and breaking of bonds to understand the DNA cleavage mechanism of FnCas9 systems which will aid in optimizing Cas9's activity.

CHAPTER 4

Deciphering Cas9 Specificity: Role of Domain Dynamics and RNA:DNA Hybrid Interactions Revealed through Machine Learning and Accelerated Molecular Simulations

4.1 Introduction

CRISPR/Cas9, a versatile genome editing tool, is widely used in exploring the functionality of genetic elements, generating genetically modified organisms, and investigating genetic disorder therapeutics in preclinical research Jinek *et al.* (2012*b*); Sternberg *et al.* (2014). *Streptococcus pyogenes* Cas9 (SpCas9) remains extensively utilized in CRISPR/Cas gene editing due to its robust activity and ease of delivery. Its ability to cleave genomic DNA sequences with partial complementarity to guide RNA (gRNA) leads to off-target genetic modifications causing genomic instability and disruption of the functionality of otherwise normal genes, limiting its widespread use for therapeutic and clinical applications Cameron *et al.* (2017*b*); Hsu *et al.* (2013); Pattanayak *et al.* (2013*b*). Cas9-induced DNA breaks at off-target sites can promote chromosomal rearrangements or translocations, potentially disrupting gene regulation or leading to aberrant fusion genes Ihry *et al.* (2018). These concerns highlight the significance of minimizing off-target effects in Cas9-mediated genome editing to ensure accurate and safe genetic modifications.

Off-target activity of CRISPR/Cas9 depends on several factors, such as the inherent specificity and quantity of the Cas9 protein, the length, composition and structure of the guide RNA (gRNA), the target cell type and its specific state, as well as the characteristics of the target site Hsu *et al.* (2013); Fu *et al.* (2013, 2014*b*); Moreno-Mateos *et al.* (2015); Zhang *et al.* (2015*b*); Duan *et al.* (2014). Numerous strategies have emerged to mitigate these effects, including Cas9 protein engineering, utilization of natural highly specific Cas9s, advanced genome editing tools like base and prime editors, optimized gRNA designs, controlled Cas9 delivery methods, and combination approaches with anti-CRISPR proteins or inhibitors. Cas9 nuclease engineering remains a prominent avenue in these efforts, reflecting the primary focus of prior studies Chen *et al.* (2020*b*); Koo *et al.* (2015); Li *et al.* (2019); Manghwar *et al.* (2020); Naeem *et al.* (2020). "High-fidelity" Cas9 variants that contain amino-acid substitutions, such as eSpCas9, Cas9-HF1, HypaCas9, $SpCas9^{2Pro}$ have been engineered through structure-guided rational approaches, while evoCas9, Sniper-Cas9 and xCas9 3.7 have been identified by directed evolution Casini *et al.* (2018*b*); Chen *et al.* (2017*a*); Kleinstiver *et al.* (2016*a*); Slaymaker *et al.* (2016*b*); Lee *et al.* (2018*b*). These variants contain amino-acid substitutions that disrupt protein-DNA/RNA-DNA interactions, or destabilize DNA-bound domains, impeding the crucial conformational change in the HNH nuclease domain. Despite enhancing mismatch discrimination, these variants often lead to a trade-off, resulting in de-

creased target DNA cleavage activity Bravo *et al.* (2022*a*). Another approach that is explored for precision optimization involves favoring homology-directed repair (HDR) over non-homologous end joining (NHEJ). HDR offers precise alterations using external templates, ensuring accuracy, targeted modifications, and reduced off-target effects compared to the error-prone NHEJ Yang *et al.* (2020). Since HDR is restricted to the S and G2 phases and relies on a repair template, its occurrence is less frequent than NHEJ Hustedt and Durocher (2017). This limitation poses a barrier for applications requiring precise sequence modifications. Therefore, improving the Cas9 specificity while preserving the target editing activity with HDR-based modification at the cleavage site is critical to exploring high-fidelity and efficient Cas9 tools.

Naturally occurring, Cas9 ortholog derived from the bacterium *Francisella novicida*, FnCas9 exhibits high specificity toward its target while demonstrating minimal affinity for off-target sites. Its cleavage activity diminishes in the presence of two base-pair mismatches, particularly at the PAM distal end indicating superior specificity compared to SpCas9 Acharya *et al.* (2019*b*). The generation of staggered ends at the cleavage site increases rates of homology-directed repair within cells, making it an apt candidate for precise genetic engineering and potential gene therapy applications Acharya *et al.* (2019*b*). Determining the structural factors responsible for the higher specificity of FnCas9 can be utilized to engineer highly specific Cas9 proteins facilitating safer and broader applications of CRISPR technology at a genome-wide level. The initial effort made to unravel these structural elements involved atomistic simulations of apo and gRNA-bound forms of SpCas9 and FnCas9 proteins over 1000 ns Panda and Ray (2022). It uncovered several structural disparities, including substantial domain deviations and increased gRNA binding affinity in FnCas9, predominantly driven by polar electrostatic interactions. To delve deeper into the structural basis behind mismatch discrimination and specificity, further investigations should focus on the ternary complexes of both Cas9s in the presence of sgRNA, target DNA, and non-target DNA—studied in native forms and with base-pair mismatches. Similar studies have been performed on SpCas9 protein which elucidated the critical role of mismatch types, their positioning, and the overall count in governing off-target DNA binding and cleavage Boyle *et al.* (2017); Doench *et al.* (2016); Jones *et al.* (2021); Zhang *et al.* (2020*b*). PAM-proximal mismatches in the gRNA-target strand (TS) seed region significantly affect substrate DNA binding and R-loop formation Boyle *et al.* (2021); Ivanov *et al.* (2020); Zhang *et al.* (2020*b*). Conversely, PAM-distal mismatches permit stable binding but often lead to the formation of catalytically impaired complexes Dagdas *et al.* (2017); Yang *et al.* (2018). The impact of base-pair mismatches on RNA:DNA heteroduplexes using molecular dynamic simulations and cryo-electron microscopy revealed the presence of base pair mismatches at specific PAM distal sites of the RNA:DNA hybrid can reduce the conformational mobility of the HNH domains, locking it in an inactive state preventing its activation toward DNA cleavagesRicci *et al.* (2019); Bravo *et al.* (2022*a*). Also, the kinetics-guided cryo-electron microscopy analysis showcased distinct conformational shifts in Cas9 structures when mismatches were present. The study revealed that the gRNA-TS heteroduplex can exist in linear conformation which is an early intermediate stage and a kinked conformation is crucial for structural rearrangement of HNH depending on the number and position of base-pair mismatches Bravo *et al.* (2022*a*). These findings emphasize the critical role of base-pair mismatches

in altering heteroduplex conformation and inhibiting Cas9 activation, elucidating the molecular mechanisms behind off-target effects. However, these studies predominantly focused on SpCas9, leaving a gap in understanding of mismatches in the paradigm of specificity. Hence, taking FnCas9 as a candidate of higher specific Cas9, in tandem with SpCas9, will enlighten the characterization of heteroduplex dynamics and elucidate the molecular determinants governing off-target binding. Therefore, studying heteroduplex dynamics with FnCas9, a candidate for higher specificity, in tandem with SpCas9, will elucidate the molecular determinants governing off-target binding.

CRISPR/Cas9 orthologs with longer and more restrictive PAMs are considered safer alternatives to SpCas9 Chen *et al.* (2017*a*); Kleinstiver *et al.* (2016*a*); Slaymaker *et al.* (2016*b*); Lee *et al.* (2016); Kostyushev *et al.* (2019); Vakulskas *et al.* (2018*b*); Müller *et al.* (2016). Ortholog derived from *Streptococcus thermophilus*, StCas9 is smaller than SpCas9 recognizes PAM- 5'-NNAGAAW-3' and 5'-NGGNG-3' for CRISPR1 and CRISPR3 loci, respectively. It demonstrates high on-target and very low off-target activity in human cells Müller *et al.* (2016); Agudelo *et al.* (2020); Zhang *et al.* (2020*c*), making it an attractive candidate for therapeutic applications. Previous studies have shown that both FnCas9 and StCas9 share higher specificity compared to SpCas9, which may be driven by unique structural factors that differentiate their interactions with target DNA. Including StCas9 in the comparison with FnCas9 can help elucidate these structural determinants more comprehensively Hu *et al.* (2018). If both FnCas9 and StCas9 exhibit similar protein factors contributing to base-pair mismatch discrimination and specificity, it would reinforce the notion that these structural factors are key to the generalizable principles governing Cas9 specificity across diverse orthologs.

In the present study, gaussian accelerated molecular dynamics (GaMD) simulations were applied to characterize structural determinants aiding base-pair mismatch discrimination and thereby contributing to the higher specificity of Cas9. Simulations were performed on fully matched RNA:DNA hybrids (on target) and mismatched hybrids (off-target) in SpCas9, FnCas9, and StCas9 systems. The results highlighted unfavorable pyrimidine-pyrimidine base pairs and preserved heteroduplex structure in FnCas9, which follows a shorter allosteric pathway for efficient information transfer, contributing to higher sensitivity against mismatches. This comparative analysis provides crucial insights into the factors contributing to precise genome editing using CRISPR/Cas9.

4.2 Materials and Methods

4.2.1 Structural Modeling

Atomistic molecular dynamics simulations have been performed using the X-ray structure of the *Streptococcus pyogenes* Cas9 in complex with RNA and DNA (PDB ID: 5F9R) Jiang *et al.* (2016), solved at 3.40 Å the resolution, which characterizes the pre-cleavage state of SpCas9. Similarly, MD simulations were conducted in the ternary form of *Francisella novicida* Cas9 in complex with RNA and DNA Hirano *et al.* (2016) (PDB ID: 5B2O), resolved at 1.7 Å. In the case of FnCas9,

the nontarget strand (ntDNA) was extended by ten nucleotides, utilizing the ntDNA strand of SpCas9 as a template. Subsequently, this template strand was mutated using Chimera Pettersen *et al.* (2004) according to the sequence in FnCas9 and aligned with the crystal structure. In the SpCas9 system, the respective lengths of the protein, sgRNA, tDNA, and ntDNA components were 1362 amino acids, 114 nucleotides (nts), 29 nucleotides, and 18 nucleotides. In contrast, for FnCas9, these components exhibited different lengths: 1622 amino acids, 92 nts, 29 nts, and 17 nts, and a Zinc ion. In the SpCas9 system, two Mg2+ ions were introduced, utilizing the Mn2+ coordinates derived from the Apo form of SpCas9 (PDB ID: 4CMP) Jinek *et al.* (2014*b*). The nucleotides on the sgRNA strand of SpCas9 and FnCas9 were mutated to target the human Emx1 gene sequence (hEMX1) (5′-GAGUCCGAGCAGAAGAAGAA-3′). Based on the X-ray structures, eight distinct model systems were established. These included configurations representing the on-target DNA sequence (referred to as "on-target") and systems with base pair mismatches at PAM-distal end of the sgRNA:tDNA heteroduplex, designated as mm20, mm16_19, and mm17–20 (as depicted in Figure 4.2A). Base pair mismatches within the sgRNA-tDNA hybrid have been introduced by substituting the nucleobases in the sgRNA strands of both the Cas9s (i.e., $3'$ – GAGUC– $5'$, Figure 4.2A), with nucleobases that do not match the DNA TS, resulting in the following base pair mismatches within the sgRNA:tDNA hybrid: G:U (at position 20), A:C (at position 19), G:U (at position 18), U:G (at position 17), and C:A (at position). All constructed model systems were minimized in Chimera Pettersen *et al.* (2004) and used as starting structures for performing simulations. These were immersed in explicit water with the addition of Na+ ions to maintain charge neutrality, creating orthorhombic periodic simulation cells, containing $\sim$250,000 atoms for the SpCas9 system and $\sim$300,000 atoms for the FnCas9 system. For a comparative analysis of allosterism, the X-ray structure of the *Streptococcus pyogenes* Cas9 in complex with RNA and DNA (PDB ID: 4UN3), solved at 2.58 Åresolution, which identifies the inactivated state of the HNH domain (i.e., the so-called “conformational checkpoint”) was also simulated following the same simulation protocol for 500 ns Anders *et al.* (2014*b*). For comparison, the X-ray crystal structure of *Streptococcus thermophilus* Cas9 (PDB ID: 6M0W) in its catalytic state with AAAGAA PAM, was considered for performing simulation following the above protocol (75ns (GaMD, and 500 ns (cMD)) Zhang *et al.* (2020*b*) . Additionally, similar base-pair mismatch systems were modeled and simulated, focusing on mismatches located at the PAM distal end: the 20th base-pair position (mm20), the 16th and 19th positions (mm16_19), and the 17th to 20th base-pair positions (mm17-20), with each system simulated for an equivalent duration.

4.2.2 Conventional MD simulations

Energy-minimized structures of the ternary complexes of SpCas9 and FnCas9 were employed as initial configurations for classical molecular dynamics (MD) simulations. These MD simulations were conducted to equilibrate the systems and prepare them as the initial states for gaussian accelerated MD (GaMD) simulations. The model systems have been equilibrated through conventional MD using the Amberff14SB Maier *et al.* (2015) force field, which includes the ff99bsc0

corrections for DNA Pérez *et al.* (2007) and the ff99bsc0+ χOL3 corrections for RNA Banás *et al.* (2010); Zgarbov *et al.* (2011). The Åqvist Aqvist (1990) model has been used for Mg2+ ions. These procedures were performed using the LEaP module of AmberTools22 HMADA *et al.* (2022). An integration time step of 2 fs was used. Hydrogen atoms were added assuming standard bond lengths and were constrained to their equilibrium position with the SHAKE algorithm. Temperature control has been performed via Langevin dynamics, Turq *et al.* (1977) with a collision frequency γ = 1. Pressure control was accomplished by coupling the system to a Berendsen barostat Berendsen (1999) at a reference pressure of 1 atm and with a relaxation time of 2 ps. The energy minimization was performed to relax water molecules and counter-ions. The protein, RNA, DNA, Mg2+, and Zn2+ were fixed with 100 kcal/molÅ^2 harmonic restraints for an initial 50,000 steps. Subsequently, a 30,000-step minimization was conducted with no restraints. The system was gradually heated from 0 to 100 K in the NVT ensemble with 50 kcal/mol restraints for 100 ps. The temperature was further increased to 200 K for the next $\sim$ 100 ps in the NPT ensemble, with restraints reduced to 25 kcal/molÅ^2+. After releasing all restraints, the system's temperature was raised to 300 K during a single 500 ps NPT simulation. Afterward, all restraints were released, and the system's temperature was raised to 300 K in a single NPT simulation of 500 ps. Post equilibration, the production run was carried out for 50 ns in the NPT ensemble using the GPU version of AMBER22 HMADA *et al.* (2022). The well-equilibrated systems have been used as a starting point for GaMD simulations.

4.2.3 Gaussian Accelerated MD simulations

The system potential energy surface (PES) is smoothed using the improved sampling technique known as "accelerated MD" (aMD), which reduces energy barriers and speeds up transitions between low-energy states by introducing a non-negative boost potential Hamelberg *et al.* (2004). In GaMD, the boost potential applied follows a Gaussian distribution, which is a more robust and novel way of accelerating simulations Miao *et al.* (2015). This enables accurate reweighting via cumulant expansion to the second order in order to acquire the canonical ensemble by reconstructing the potential energy surface's original shape. Therefore, the same low-energy physical states are sampled in the GaMD simulations even with biasing potential. As a result, reweighting can enable the quantitative recovery of conformational distributions, whereas results that are not reweighted, such as those shown here, capture the low-energy physical states and also offer a helpful semiquantitative ranking of their probability. This technique helps to describe the canonical distribution for a large biomolecular system, such as the CRISPR/Cas9 Palermo (2019*b*); Ricci *et al.* (2019) biomolecular system, such as CRISPR/Cas9 Palermo (2019*b*); Ricci *et al.* (2019), according to available experimental data. GaMD extends the use of aMD to large biological systems, for which it has been challenging to attain the canonical ensemble, given the large statistical noise Miao and McCammon (2016). Considering a system with N atoms at positions $\vec{r} = \{\vec{r_1}, ...\vec{r_N}\}$ when the system potential $V(\vec{r})$ is lower than a threshold energy E, the energy surface is modified

by adding a boost potential as

$$V^*(\vec{r}) = V(\vec{r}) + \triangle V(\vec{r}) \tag{4.1}$$

$$\triangle V(\vec{r}) = 1/2k(E - V(\vec{r}))^2 \tag{4.2}$$

where k is the harmonic force constant. The two adjustable parameters E and k are automatically determined by applying the following three criteria. First, for any two arbitrary potential values $V_1(\vec{r})$ and $V_2(\vec{r})$ found on the original energy surface, if $V_1(\vec{r}) < V_2(\vec{r}), \triangle V$ should be a monotonic function that does not change the relative order of the biased potential values, i.e., $V_1^*(\vec{r}) < V_2^*(\vec{r})$. Second, if $V_1(\vec{r}) < V_2(\vec{r})$, the potential difference observed on the smoothened energy, i.e., $V_2^*(\vec{r}) - V_1^*(\vec{r}) < V_2(\vec{r}) < V_2(\vec{r})$. By combining the first two criteria and putting in the formula of $V^*(\vec{r})$ and $\triangle V(\vec{r}$, we obtain the following result.

$$V_{max} \leq E \leq V_{min} + 1/k \tag{4.3}$$

where V_{min} and V_{max} are the system minimum and maximum potential energies. To ensure that eq4.4 is valid, k has to satisfy $k \leq 1/V_{max} - V_{min}$. By defining $k \equiv k0(1/V_{max}) - V_{min}$, then $0 < k \leq 1$. Third, the standard deviation of $\triangle V$ needs to be small enough (i.e., narrow distribution) to ensure accurate reweighting using cumulant expansion to the second order: $\sigma_{\triangle V} = k(E - V_{avg})\sigma_V \leq \sigma_0$, where V_{avg} and σV are the average and standard deviation of the system potential energies, $\sigma_{\triangle V}$ the standard deviation of $\triangle V$, and σ_0 is a user-specified upper limit (e.g., 10 kBT) for accurate reweighting. When E is set to the lower bound, E = V_{max}, according to eq4.4, k_0 can be calculated as

$$k_0 = min(1.0, k_0') = min\left(1.0, \frac{\sigma_0}{\sigma_V}\frac{V_{max} - V_{min}}{V_{max} - Vavg}\right) \tag{4.4}$$

Alternatively, when the threshold energy E is set to its upper bound E = $V_{min} + 1/k$, k_0 is

$$k_0 = k_0'' \equiv \left(1 - \frac{\sigma_0}{\sigma_V}\right)\frac{V_{max} - V_{min}}{V_{avg} - Vmin} \tag{4.5}$$

if k_0'' is calculated between 0 and 1. Otherwise, k_0 is calculated using eq4 instead of being set to 1. The system threshold energy was set to E = V_{max} for all GaMD simulations. The boost potential has been applied in a dual-boost scheme, in which two acceleration potentials are applied simultaneously to the system: (i) the torsional terms only and (ii) across the entire potential. A time step of 2 fs was used. The maximum, minimum, average, and standard deviation values of the system potential (V_{max}, V_{min}, V_{avg}, and σ_V) were obtained from ∼50 ns cMD with no boost potential. Each GaMD simulation proceeded with a ∼25 ns run, in which the boost potential was updated every 1.6 ns, thus reaching equilibrium values. Finally, ∼ 500 ns of GaMD simulations were performed in the NPT ensemble. For each system, ∼500 ns (GaMD production) + 25 ns (GaMD equilibration) + 50 ns (pre-equilibration conventional MD) was carried out. Since GaMD has been applied on 8 model systems, including on and off-target DNAs, a total of ∼4.6 μs (i.e., ∼500 ns of production + 75 ns of equilibration × 8 systems) of simulations has been produced. The MD trajectory snapshot was saved after every 1 ns. Only the stable part of the trajectory was used for simulation analysis (300-500 ns). These simulations have been performed with the GPU

version of AMBER22.

4.2.4 MD Simulation Trajectory Analysis

Geometrical properties crucial for understanding and comparing the system's behavior were assessed by analyzing the equilibrium MD trajectories. These properties involved the root mean square deviation (RMSD) of protein backbone atoms (Å) (Figure 4.1A and 4.1B) and RNA:DNA hybrid (Figure 4.1C and 4.1D), the solvent-accessible surface area (SASA) of Cas9 (Figure 4.1G-H), as well as the radius of gyration of Cas9 (Figure 4.1E and 4.1F). The residue-wise root mean square fluctuations were also computed for both the Cas9s in their on-target and off-target forms. In addition, the analysis involved quantifying tight contacts within a 4 Åradius between the sgRNA and tDNA within the heteroduplex. Furthermore, native and non-native contacts formed by the heteroduplex with neighboring protein domains, including HNH, RuvC, and REC, were examined. To gain insight into the orientational and angular dynamics, the angles between specific atom pairs were calculated for both SpCas9 and FnCas9 (on-target and off-target systems) during the stable portion of the trajectory. This included the angle formed by DC:1508, DC:1494, and DA:1527 atoms in SpCas9, and DC3:1746, DC:1732, and DG3:1765 atoms in FnCas9. These analyses provided essential data to elucidate the structural characteristics and interactions within the systems studied. All the structures were rendered using PyMOL Schrodinger and DeLano (2020*b*).

4.2.5 Analysis of the RNA:DNA base-pair complementarity

The sgRNA:tDNA structural dynamics were analyzed during the GaMD production runs using the 3DNA code Lu and Olson (2003). The nastruct module of the cpptraj Roe and Cheatham III (2013) present in AmberTools22 Case *et al.* (2023) was utilized to calculate RNA:DNA geometrical parameters from the trajectory Babcock *et al.* (1994); Olson *et al.* (2001). Specifically, rotational and translational parameters within each base pair, characterizing base pair complementarity (e.g., shear, stretch, stagger, buckle, propeller, and opening), were computed for the stable part of the trajectory (last 200 ns). The number of native contacts (within 4 Å) formed between the ntDNA-tDNA (Figure 4.2A-B) and sgRNA-tDNA strands (Figure 4.2C) was computed for all the forms of SpCas9 and FnCas9 for the stable part of the trajectory. The 3DNA tool calculates key base-pairing parameters, such as hydrogen bond patterns, base-pair distances, and angular distortions (e.g., shear, stretch, stagger, propeller, opening, and buckle), and compares these values to the idealized geometries of Watson-Crick pairs. If the observed interactions deviate significantly from the standard A-T or C-G hydrogen bonding and geometrical norms, 3DNA classifies them as non-Watson-Crick pairs. These deviations often arise due to altered hydrogen bonding, base stacking, or structural distortions induced by protein-DNA interactions or unusual DNA conformations. Based on this, the loss of base pairing (Figure 4.2D) and the count of non-canonical base pairs (Figure 4.2C) formed within the RNA:DNA hybrid in mismatch systems for both Cas9 proteins were also assessed, utilizing the 3DNA code.

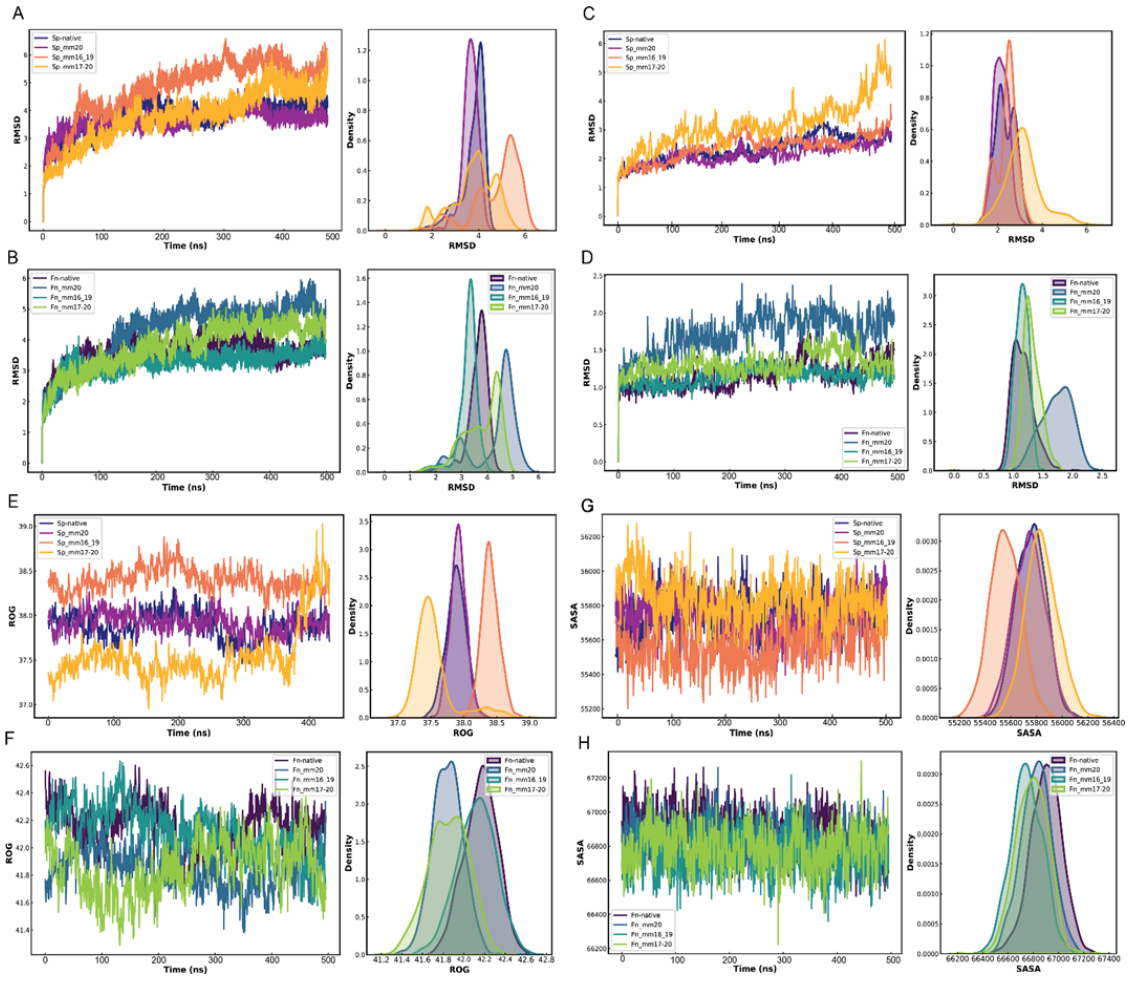


Figure 4.1: (A) RMSD of SpCas9 backbone atoms for native and off-target forms with RMSD probability distributions; (B) RMSD of FnCas9 backbone atoms for native and off-target forms with probability distributions; (C-D) RMSD of RNA hybrid atoms for SpCas9 and FnCas9, respectively, with probability distributions; (E-F) Radius of gyration of SpCas9 and FnCas9 protein atoms with probability distributions; (G-H) Solvent Accessible Surface Area (SASA) of SpCas9 and FnCas9 with probability distributions for native and off-target forms.

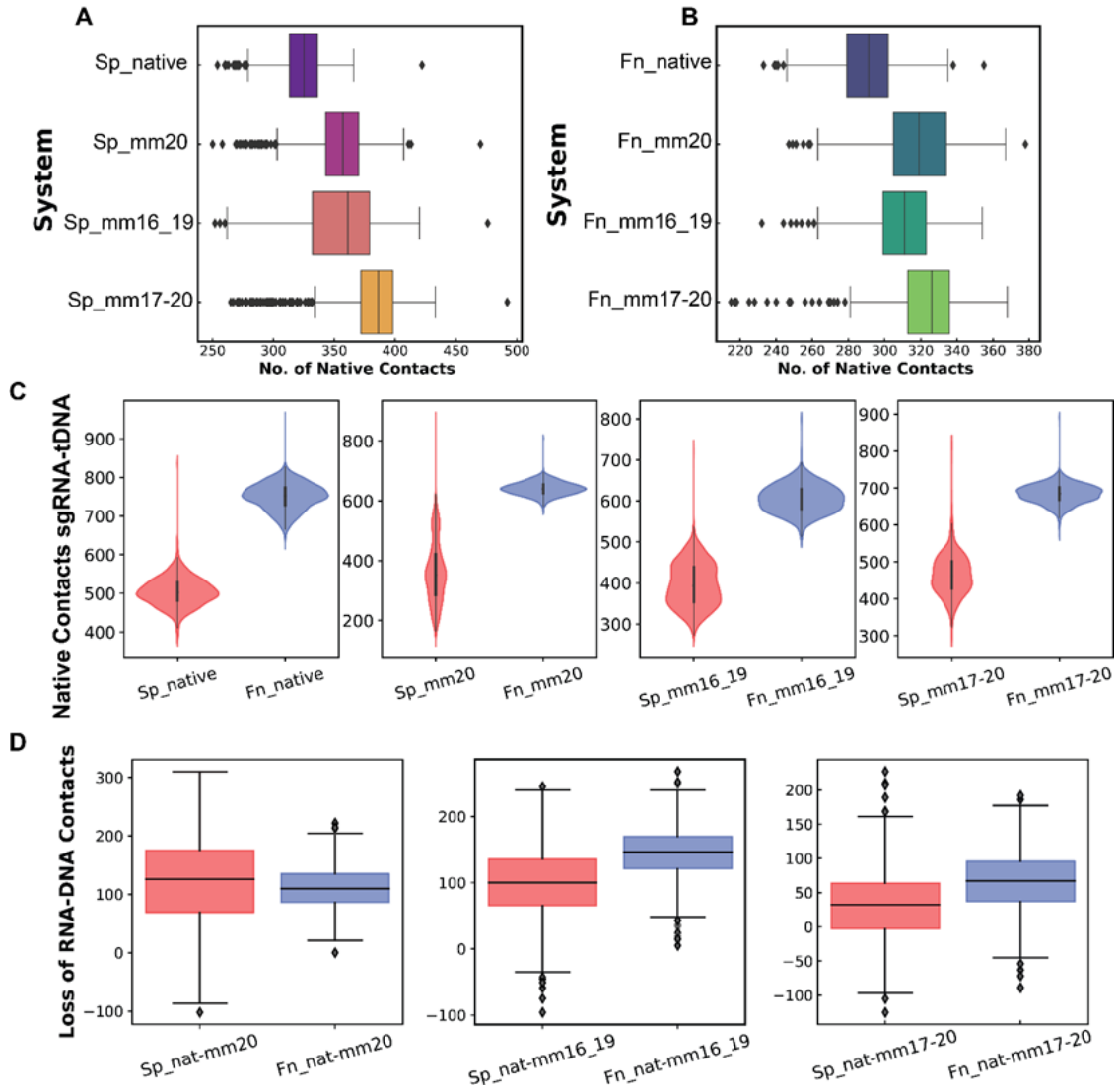


Figure 4.2: Distribution (as boxplot) of the number of native contacts formed between ntDNA:tDNA in all the forms of (A). SpCas9, (B). FnCas9. (C). Pair-wise distribution (as violin plot) of the number of native contacts formed between sgRNA:tDNA in all the forms SpCas9 and FnCas9. (D) Pair-wise distribution (as boxplot) of the loss of RNA:DNA contacts formed in all the off-target forms SpCas9 and FnCas9.

4.2.6 Principal component analysis

To investigate the dominant motions and conformational distributions of RNA:DNA heteroduplex and Cas9 domains (HNH, RuvC, REC1, REC2, and REC3 domains) in native form and in the presence of base-pair mismatches, principal component analysis (PCA) was performed Wan *et al.* (2013, 2015). This approach involves the calculation of a covariance matrix, followed by a linear transformation that projects the macromolecular dynamics into the low-dimensional space. The covariance matrix was built and diagonalized using the Pytraj module Roe and Cheatham (2013). In the case of SpCas9, the analysis was performed using the cartesian coordinates for the sgRNA-tDNA heteroduplex (residues 1365-1384, and 1490-1508), RuvC (residues 1-56, 718-765, and 918-1099), HNH (residues 780-906), REC1 (residues 94-167, and 308-497), REC2 (residues 168-307) and REC3 (residues 498-713) domains. Similarly, for FnCas9, these calculations were performed for analogous structural components, including the sgRNA-tDNA heteroduplex (residues 1365-1384 and 1490-1508), RuvC (residues 1-51, 858-899 and 1088-1224), HNH (residues 932-1070), REC1 (residues 83-112, and 307-459), REC2 (residues 112-307) and REC3 (residues 459-858) domains. The analysis used all frames from all equilibrated simulations in on-target and off-target forms of both Cas9s.

4.2.7 Linear Interaction Energy

The electrostatic and Van der Waals interaction energies between gRNA and tDNA were calculated using the linear interaction energy model (20 base pair region) Brandsdal *et al.* (2003); Åqvist *et al.* (2002) for on-target and off-target forms of both the Cas9s. The linear interaction energy of binding is linearly proportional to the differences in van der Waals (modeled using Coulomb potential energy functions) and electrostatic interactions (modeled using the Lennard-Jones function) involving the ligand in its bound and free state in the solvent. These values are scaled by LIE parameters α and β, respectively. The cut-off distance was set at 12 Å.

$$\triangle E_{LIE} = \alpha \left(E^{vdW}_{bound} - E^{vdW}_{free} \right) + \beta \left(E^{ele}_{bound} - E^{ele}_{free} \right) \tag{4.6}$$

4.2.8 Markov State Modelling of Heteroduplex

Markov state models (MSMs) are a powerful approach to extracting long-lived metastable macrostates, their equilibrium occupancy populations, and the transition among these states. It provides an interpretable and predictive model of the system thermodynamics and kinetics. In this study, MSM was performed to determine the metastable states of the heteroduplex in the on-target and off-target forms of the SpCas9 and FnCas9 proteins and to study the transition between these states. To conduct this analysis, cartesian coordinates of 20 bp long sgRNA-tDNA heteruduplex of both Cas9 proteins were considered. All the trajectory frames generated for 500ns long trajectory were taken. MSMs were constructed for each of the eight sgRNA-tDNA heteroduplexes, 8 × 500 ns simulation

trajectories following protoco (shown in Figure 4.4): (i) Trajectory featurization, (ii) Dimensionality reduction, (iii) Discretization of Trajectories, and (iv) MSM estimation and validation (v) PCCA Deuflhard and Weber (2005) and Transition Probability Analysis. Calculations were performed using the PyEMMA software package Scherer *et al.* (2015). Trajectories representing the cartesian coordinates of the 20 bp long sgRNA-tDNA heteroduplexes over time were featurized using the MDTraj Python libraries McGibbon *et al.* (2015). Specifically, the intermolecular pairwise distances (d(i,j)) between the center of mass of the four bases at the PAM distal end (i.e., positions 20th, 19th, 18th, and 17th base-pair positions) were utilized as input features for constructing the Markov State Model (MSM). The featurized trajectories were generated for all forms of both Cas9s and concatenated to create two featurized trajectories, each containing features for both the on-target and off-target forms of SpCas9 and FnCas9, respectively. The featurized trajectories obtained for SpCas9 and FnCas9 were then projected into a low-dimensional space in preparation for microstate clustering using two-component principal component analysis and time-independent Component Analysis (tICA), which maximally preserves the kinetic variance in the data. PCA results were considered for further analysis since the active-state fraction was less than one for the tICA method. Subsequently, the trajectories were clustered into 200 cluster centers (macrostates) by k-means clustering, as implemented in PyEMMAScherer *et al.* (2015). A 200 microstate MSM was built at varying lag times. Finally, a lag time of around 300 steps (that is, 6 ns) was chosen as the minimum value of lag time at which the implied time scale (ITS) becomes flatter as a signature of the Markovian nature of the model. The optimum number of macrostates (five for SpCas9 and four for FnCas9) for every trajectory was determined based on the relaxation timescale separation. The 200 microstates were further coarse-grained into five macrostates for SpCas9 and four macrostates for FnCas9 using the Perron Cluster Cluster Analysis (PCCA+) method Deuflhard and Weber (2005); Röblitz and Weber (2013). The construction of these macrostate MSM for both Cas9 systems was validated by performing the Chapman-Kolmogorov test with a Bayesian error estimation model Noé *et al.* (2009). A perfect agreement between estimated and predicted models at a higher lag time 1 step suggests a valid MSM model. Kinetic models were generated based on the calculated transition-state probabilities for every state.

4.2.9 Uniform Manifold Approximation and Projection Clustering of DNA Structural Properties and Free-Energy Landscape Analysis

Uniform Manifold Approximation and Projection (UMAP), a non-linear dimension reduction technique, was used to project the structural parameters of RNA:DNA hybrid and proteins into a two-dimensional space Trozzi *et al.* (2021). Two crucial hyperparameters, namely the maximum distance and the minimum number of neighbors were used to maintain the balance between local and global structure in the final projection. The UMAP projections were fine-tuned by adjusting the 'min_dist' (0-0.5) and 'n_neighbours' (0-200). The Manhattan distance is the metric to assess the distances between points in both the high- and low-dimensional manifolds. The silhouette score was calculated to compute the separation between and within the clusters to compare the

distinctness of the cluster for all systems Shutaywi and Kachouie (2021). These computations were conducted using the 'umap-learn' Python package. After hyper-tuning, the optimized values of 'min_dist' and 'n_neighbors' were selected for the final UMAP projections based on the silhouette score. Silhouette score was calculated to compute the separation between and within the clusters to compare the distinctness of the cluster for all systems [76]. It ranges from -1 to +1, evaluating how well each data point fits into its assigned cluster. A score close to +1 indicates the data point is well-clustered, while a negative score suggests it may have been assigned to the wrong cluster. The structural properties used for this analysis:

RNA:DNA hybrid structural properties: The intra-base pair rotational (propeller, buckle, and opening), translational parameters (shear, stagger, and stretch) of the 20th, 19th, 18th, and 17th base pairs, the number of tight contacts (within 4Å) between sgRNA-tDNA and the linear interaction energy between both the sgRNA and tDNA strands. These features were computed for the stable part of the trajectory (300-500ns) for all on-target and off-target forms of SpCas9 and FnCas9.

Protein structural properties: The protein- RMSD (backbone atoms), the radius of gyration (ROG), solvent accessible surface area (SASA), angle between atoms DC-1508, DC-1494, and DA-1527) in SpCas9, and the angle between atoms DC3-1746, DC-1732, and DG3-1765) in FnCas9 and the number of native and non-native contacts between RNA:DNA hybrid and HNH, RuvC and REC1, REC2 and REC3 domains were calculated for stable part of trajectory for all on-target and off-target systems of SpCas9 and FnCas9.

The analysis was conducted in three modalities: (a) RNA:DNA hybrid structural properties alone, (b) protein structural properties alone, and (c) RNA:DNA hybrid and protein structural properties.

4.2.10 Machine Learning modeling to classify heteroduplex states in SpCas9 and FnCas9 systems

MSM was performed to determine the metastable states of the heteroduplex in on-target and off-target forms of the SpCas9 and FnCas9 proteins. Machine learning methods were employed to identify the structural characteristics that aid in classifying the macrostates obtained from MSM, highlighting the essential features contributing to their distinction. The workflow used for machine learning prediction is shown in 4.4-Metastable State Prediction.

Training and Testing Dataset

The following structural descriptors were used to train the machine learning algorithm (ML).

RNA:DNA hybrid structural properties: The intra-base pair rotational (propeller, buckle, and opening), translational parameters (shear, stagger, and stretch) of the 20th, 19th, 18th, and 17th

base pairs, the number of tight contacts (within 4 Å) between sgRNA-tDNA and the linear interaction energy between both the sgRNA and tDNA strands. These features were computed for the complete trajectory for all on-target and off-target forms of SpCas9 and FnCas9.

Protein structural properties: The protein-RMSD (backbone atoms), the radius of gyration (ROG), solvent accessible surface area (SASA), the angle between atoms DC-1508, DC-1494, and DA-1527) in SpCas9, and the angle between atoms DC3-1746, DC-1732, and DG3-1765) in FnCas9 and the number of native and non-native contacts between RNA:DNA hybrid and HNH, RuvC and REC1, REC2 and REC3 domains were calculated for the complete trajectory for all on-target and off-target systems of SpCas9 and FnCas9.

Metastable State Determination: The Markov State Model (MSM) framework was applied to identify the distinct metastable states of the heteroduplex, which were used as labels for the classification task. These labels were also obtained for all the frames present in on-target and off-target forms of both the Cas9s.

Feature Selection Framework

This study used various feature selection and ranking techniques to identify the critical structural determinants responsible for distinguishing metastable states within the SpCas9 and FnCas9 protein systems. The correlation between all protein and hybrid-based features was visualised using a heatmap shown in Figure 4.3(using Pearson correlation coefficient). Due to the non-linear nature of the relationships between features and the target variable, non-linear machine learning models were employed to investigate their relationships. The dataset for SpCas9 and FnCas9 consisted of 57074 frames and 52080 frames, respectively, where a total of 42 features were extracted for both Cas9 variants. Two primary categories of feature selection methods were utilized: Wrapper and Embedded. Wrapper methods assess feature subsets by feeding them into a model, while Embedded methods identify the most valuable features during the model creation process. Specifically, two Wrapper-based feature selection methods, BorutaSHAP (based on XG Boost), which combines the Boruta feature selection algorithm with Shapley values, and Recursive Feature Elimination (RFE), a technique that iteratively removes features and constructs models with the remaining ones were implemented Gramegna and Giudici (2022); Zeng *et al.* (2009). Additionally, Extra-Tree Classifier, a tree-based embedded feature selection method, was implemented to calculate feature importances while training an ensemble of decision trees.

The Extra-Tree Classifier (Extremely Randomized Trees) was chosen as our primary embedded feature selection and classification model due to its ability to handle complex, high-dimensional datasets and its computational efficiency. Unlike traditional decision tree algorithms, Extra-Trees introduce additional randomness during the tree building process, not only by bootstrapping the data but also by randomly selecting cut-points for features Arya *et al.* (2022). This increased randomness contributes to the creation of a more diverse ensemble of trees, which can lead to improved generalization performance and a more robust estimation of feature importances Kramer

and Kramer (2016). Extensive benchmarking has shown that Extra-Trees often outperform or match other ensemble methods like Random Forests, especially on classification tasks, while also being faster to train due to their randomized split selection. Importantly, Extra-Trees provide robust estimates of feature importance, making them particularly useful for embedded feature selection in high-dimensional biological dataset Arya *et al.* (2022). To further validate and refine our feature selection, we utilized BorutaSHAP, a wrapper method that combines the Boruta all-relevant feature selection algorithm with the interpretability of SHapley Additive exPlanations (SHAP) values Moen and Johnston (2023). Boruta employs a unique strategy of comparing the importance of original features against the importance of shadow features (randomly permuted copies of the original features) to determine statistically significant relevant features Kursa and Rudnicki (2010) . By integrating SHAP values, BorutaSHAP provides a more nuanced understanding of feature importance, indicating not only which features are important but also their direction of effect on the model's output Lundberg and Lee (2017). To further corroborate our feature selection results, we employed Recursive Feature Elimination (RFE), a widely used wrapper method that iteratively removes the least important features based on model performance until the optimal subset is identified. RFE works by fitting a model (such as a tree-based classifier) and ranking features according to their importance, recursively eliminating the least significant ones at each step Guyon *et al.* (2002). This process helps in identifying the minimal set of features that contribute most to predictive accuracy, thereby reducing dimensionality and potential noise in the dataset.

All structural features extracted from the trajectories of on-target and off-target forms of both the Cas9s were input into these three feature selection techniques: BorutaSHAP, Recursive Feature Elimination, and Extra-Tree Classifier Gramegna and Giudici (2022); Zeng *et al.* (2009). The selected features from these methods were used for training and evaluating the machine learning models. Machine learning predictions were performed using three distinct modalities, using: (a) RNA:DNA hybrid structural properties (28) alone, (b) protein structural properties (14) alone, and (c) both RNA:DNA hybrid and protein structural properties (42). Feature selection was independently applied for each approach.

Model Development & Assessment

The dataset was randomly partitioned into training and test sets, maintaining an 80:20 ratio. The input features comprised RNA:DNA hybrid-specific attributes, protein-specific attributes, as well as a combination of both hybrid and protein features. The output variables for these input features were heteroduplex metastable states obtained from MSM for SpCas9 and FnCas9. To ensure the robustness of our feature set, feature selection techniques were employed using Wrapper and Embedded methods. The classification of heteroduplex states was performed with selected features using several different machine learning models (Random Forest, gradient-boosting, ADA-Boost, XG Boost, and Extra-Tree Classifier). The models were trained using ten-fold stratified cross-validation to ensure that each fold maintained the proportion of classes present in the dataset, which is crucial for avoiding bias in model evaluation. This approach helps in providing a more

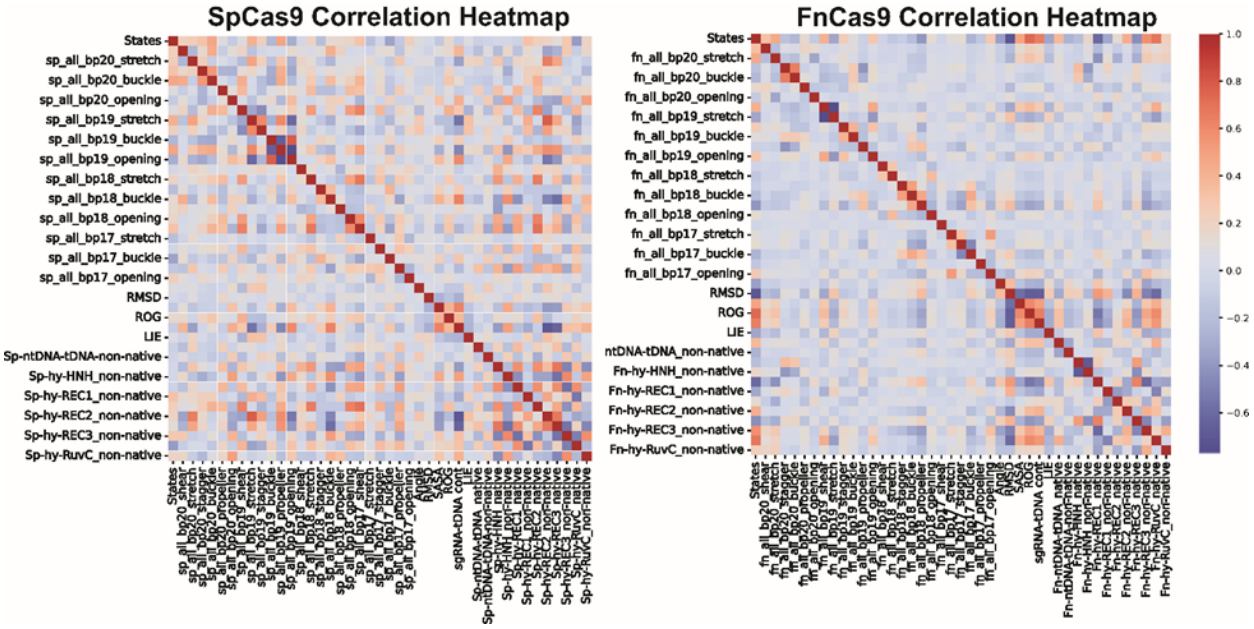


Figure 4.3: Heatmap showing correlation among the features found in SpCas9 and FnCas9

reliable estimate of model performance, particularly when dealing with imbalanced datasets. Hyperparameter tuning was carried out on selected features and was achieved using Grid Search CV. The models were evaluated using the evaluation metrics Accuracy, Precision, Recall, and F1-score. The accuracy of the model was defined as

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

where TP stands for true positive, FP for false positive, FN for false negative, and TN for true negative. The recall is the measure of how many actual positives the model can capture, while the precision is how many of the predicted positives are correct.

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

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

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

$$F1score = \frac{2 * Recall * Precision}{Recall + Precision}$$

The F1-score is a measure of a model's accuracy, considering both precision and recall, with values ranging from 0 to 1. It was used to compare the model's performance in different modalities. All models evaluated were implemented using the scikit-learn package in Python Kramer and Kramer (2016). To identify the optimal machine learning model, the F1 score was used as a selection criterion due to the inherent nature of the predictive task, which involved the classification

of outputs with multiclass labels.

4.2.11 MM/GBSA Calculations

The Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method Genheden and Ryde (2015) is a popular technique for determining binding free energies between proteins and ligands. gmx_MMPBSA Valdés-Tresanco *et al.* (2021) was used to calculate the binding free energy of the nuclease (NUC) and recognition (REC) lobes with RNA:DNA hybrid. Both the lobes were used as receptors and heteroduplex was used as the ligand. The GB (OBC) model (igb=2) was used to calculate the solvation energy. A total of 1000 snapshots corresponding to every MSM state obtained for SpCas9 and FnCas9 were selected to calculate the enthalpy contribution using MM/GBSA. The single trajectory approach was used to calculate the binding free energy. Molecular mechanics potential energy (electrostatic and van der Waals energies), the free energy of solvation (polar and nonpolar solvation energies), and the energetic contribution of each residue and nucleotide to total binding energy were obtained for all the states of SpCas9 and FnCas9 proteins. The gas-phase energies calculated using the Molecular Mechanics force field are a sum of van der Waals (VDWAALS) and electrostatic (EEL) energies, whereas the solvation energy includes a polar and a non-polar component. The polar component is calculated using an implicit Solvent Model, i.e. Generalized Born (GB), (EGB), and the non-polar component depends on the solvent accessible surface area (ESURF). Free energies of solvation are estimated by applying Poisson-Boltzmann (PB) calculations for the electrostatic contribution and a surface-area-dependent term for the non-electrostatic contribution to solvation.

4.2.12 Shortest Residue Interaction Pathway Identification

Residue interaction networks (RINs) are used to represent residue contacts in protein structures. The RING 3.0 server calculates high-quality and reliable non-covalent molecular interactions based on geometrical parameters Clementel *et al.* (2022). RINs spanning the REC3-REC2-HNH domains of both Cas9 proteins were calculated using this tool. These networks are depicted as unweighted graphs, where distinct distance thresholds were used for specific interactions: 4.5 Åfor H-bonds, 5 Åfor Ionic interactions, 7 Åfor π-cation and π-π stacking, 3 Åfor disulphide interactions, and 0.8 Åfor Van der Waals interactions. Residue-residue contact calculations were based on C-alpha atoms. In the case of SpCas9, Asn773/Arg692 in the REC3 domain served as the recognition site (source), while His773/His840 in the HNH domain was designated as the catalytic site (sink) in checkpoint/pre-cleavage states. For FnCas9, the source and sink residues were Lys840 and Asn995 in the REC3 and HNH domains, respectively. Comparative analyses involved investigating the interaction pathways within REC3-REC2-HNH domains of SpCas9 in both the checkpoint state (PDB ID: 4UN3) and the pre-cleavage state (PDB ID: 5F9R), contrasting these with the FnCas9 ternary state. Breadth-First Search (BFS) algorithm of NetworkX Python package Hagberg *et al.* (2008) was employed to find the shortest pathway between the specified source

and sink residues in both Cas9 proteins. Fifty frames comprising the coordinates of the REC3, REC2, and HNH domains were extracted from a stable part of the trajectory and given as input to the RING 3.0 server. The resulting edge and node information from the RING 3.0 output was used to determine the shortest pathway.

4.2.13 Statistical test

A non-parametric Mann-Whitney U test was used to compare the first two components of the Principal Component Analysis (PCA) generated using the coordinates of the RNA:DNA hybrid, RuvC, HNH, REC1, REC2, and REC3 domains for both on-target and off-target forms of SpCas9 and FnCas9. The aim of this analysis was to assess whether these components exhibited similar or different distributions among all the forms of SpCas9 and FnCas9 systems. Furthermore, the distribution of binding energy between the REC lobe with the RNA:DNA hybrid and the NUC lobe with RNA:DNA hybrid in the SpCas9 and FnCas9 systems was also examined using the Mann-Whitney U test. The comparison of REC-Hybrid and NUC-Hybrid binding energy distributions between SpCas9 and FnCas9 was found to be statistically significant (p-value$<$0.05), indicating distinctions in the distributions.

4.3 Results and Discussion

To decipher the factors behind specificity and to determine conformational changes induced by base-pair mismatches, all-atom GaMD simulations were performed for both Cas9 proteins. This study involved simulating the CRISPR/Cas9 complex in two scenarios: first, with a fully matched RNA:DNA hybrid (considered as the native on-target system), and second, with base pair mismatches introduced at positions 20–16 of the RNA:DNA hybrid (as depicted in Figure 4.5A). This led to the generation of distinct models denoted as mm20 (mismatch at 20th bp position), mm16_19 (mismatch at 16th and 19th bp position), and mm17–20 (mismatches from 17-20 bp positions). Each GaMD simulation in this study was conducted over a duration of 500 ns, allowing an extensive exploration of conformational space.

4.3.1 Diverse non-canonical base-pairing in the RNA:DNA hybrid enables mismatch tolerance

Previous studies have highlighted the importance of heteroduplex dynamics in understanding CRISPR-Cas9 behavior. Distal mismatches of PAM were reported to have impacted the heteroduplex, inhibiting the conformational mobility of the HNH domain Ricci *et al.* (2019). On the contrary, other observations have indicated that the mismatches embedded within the RNA:DNA hybrid had minimal impact on HNH dynamics, allowing activation and subsequent DNA cleavage Mitchell *et al.* (2020). Crystallographic structures of Cas9 bound to bonafide off-target substrates,

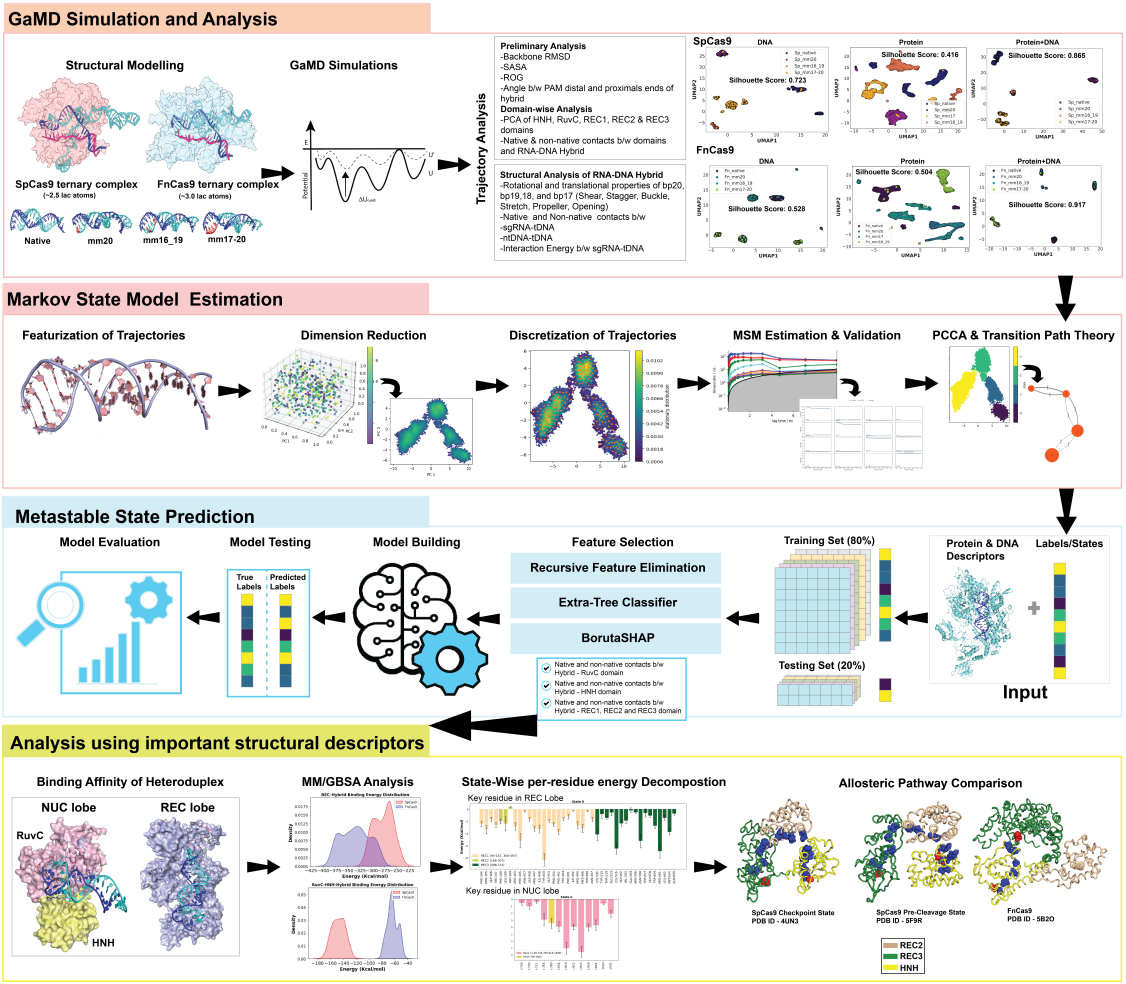


Figure 4.4: Schematic diagram: Molecular Dynamic Simulation Setup. (1) On-target and off-target structural models used for performing the simulations. (2) Post-simulation trajectory analysis conducted for analysing the structural changes in Cas9 protein and RNA:DNA heteroduplex. Construction of the Markov state model (MSM). (1) Intra-base pair contacts at the distal end of the PAM were used for feature selection. (2) Identification of reduced key features using PCA dimension reduction method. (3) Clustering similar conformations in the trajectory using K-means. (4) MSM estimation and validation using the CK test. (5) The Transition probability calculation for every state is visualized as a graph, where the nodes represent each MSM state, and the thickness of the directed edges is the transition probability from two states. Metastable State Prediction. (1) RNA:DNA hybrid and Protein descriptors were extracted from trajectories to predict heteroduplex states (given by MSM). (2) Split the data into training (80%) and testing dataset (20%). (3) Feature Selection used for improving the prediction: RFE, BorutaSHAP, and Extra-Tree Classifier. (4) Model Building and Model Testing. (5) Evaluating the model's performance based on Accuracy, Precision, F1-score and Recall. Analysis using structural descriptors. (1) Calculation of the binding affinity of the REC and NUC lobe with the heteroduplex using the MM/GBSA method. (2) Density distribution of binding energies in SpCas9 and FnCas9 systems. (3) Per-residue energy decomposition analysis to find key residues involved in Hybrid and REC/NUC lobe binding in every state (X-axis: Residues; Y-axis: Energy kcal/mol). (4) Comparison of the allosteric pathway among the precleavage and checkpoint state of SPCas9 with the precleavage state of FnCas9.

highlighted the significance of non-canonical base-pairing in mismatch tolerance within SpCas9 off-target complexes Pacesa *et al.* (2022). Furthermore, cryo-electron microscopy study demonstrated distinct conformations of the RNA:DNA duplex, linear versus kinked, associated with specific angular deviations (25 degrees vs. 75 degrees) and their direct link to Cas9 activation for DNA cleavage Bravo *et al.* (2022*a*). These findings provide insights into the structural changes induced by mismatches at the PAM-distal end suggesting that evaluating non-canonical base-pairs and the angles between PAM proximal and distal ends is pivotal in unraveling the intricate dynamics of the heteroduplex and its consequential effects on Cas9's activation and cleavage mechanisms.

To explore the change in conformational dynamics of the heteroduplex in the presence of base pair mismatches within the RNA:DNA hybrid, PCA was performed for all systems of Cas9 using the coordinates of the RNA:DNA hybrid from the stable part of the trajectory (300-500ns) as shown in Figure 1. This analysis revealed higher structural transitions in the SpCas9 heteroduplex as the PCs depicted a multimodal distribution with more than one peak (Figure 4.5B), indicating a more flexible structure. In contrast, FnCas9 heteroduplexes showed similar conformational variability in all on-target and off-target forms, suggesting a single unique conformation. Differences in heteroduplex conformational distributions between on- and off-target systems along PC1 and PC2 were assessed by performing the Mann-Whitney U tests. The native form of SpCas9 exhibited statistically significant differences in the conformational distribution compared to Sp_mm20 (p-value: 0.002) and Sp_mm16_19 (p-value: 0.010) along PC1. Similarly, the heteroduplex in the native form of FnCas9 displayed statistically significant differences in conformational distribution compared to Fn_mm20 (along PC1) and Fn_mm16_19 (along PC2). It can be inferred from the PCA analysis that the 20 bp long heteroduplex underwent less conformational deviations in FnCas9 as compared to SpCas9 (on-target and off-target forms) suggesting the structural rigidity of FnCas9 heteroduplexes even in the presence of base-pair mismatches.

Next, to understand the role of non-canonical base pairing in off-target tolerance, the number and type of non-canonical base pairs observed in SpCas9 and FnCas9 were computed for the stable part of the trajectory (Figure 4.5C). The presence of PAM-distal base-pair mismatches also resulted in the loss of Watson-Crick base pairs, thereby disrupting the structure of the heteroduplex. Hence, the loss of base pairs was also analysed to gain insights into the variations in mismatch tolerance observed between SpCas9 and FnCas9 off-target systems. In the case of SpCas9, a wide variety of non-canonical base pairs (40 types) were observed, as illustrated in Figure 4.2C, which agrees with the observation made by Jinek's group Pacesa *et al.* (2022). The most common non-canonical base pairs included G-T, C-G, A-C, C-T, A-T, and others. In contrast, FnCas9 exhibited a less diverse range of non-canonical base-pairs (16 types) in the off-target systems, with common C-T, A-T, and U-T base-pairs occurrences. Interestingly, a few base-pair types, notably the U-T base-pair, were exclusive to the systems with four mismatches in the FnCas9 (mm17-20 system). The pyrimidine-pyrimidine base pairs (C-T / U-T) known for their weaker stacking and reduced stability were more prevalent in the FnCas9 off-target systems, potentially influencing its binding to the RNA:DNA hybrid and the cleavage mechanism. The angular deviations between the PAM-distal and PAM-proximal ends of both Cas9 proteins in the presence of base pair mismatches for

the stable portion of the trajectory exhibited a range of 100°-160° in FnCas9, which was larger than the corresponding range of 60°-100° observed in SpCas9 (shown in Figure 4.5D). FnCas9 displayed a broader range of angles between the PAM-distal and proximal ends, especially in the case of a four mismatch-containing system, indicating a distinct activation mechanism and greater conformational variability compared to the SpCas9's heteroduplex.

Studies have reported structural distortions induced by off-target substrates. Therefore, the critical geometrical descriptors of base pair complementarity, including shear, stagger, propeller, opening, and stretch of the last four bases, were examined to understand the extent of nucleic acid distortions promoted by base pair mismatches in the Sp/FnCas9 systems. These parameters were extracted for the mutated base pairs at PAM-distal ends and compared for off-target bound complexes of Sp/FnCas9 systems. The dimensionality reduction of this data facilitated a more intuitive exploration of structural variations induced by these mismatches (detailed description in Section 3.9). The silhouette score was calculated to assess the observed clusters for all three cases. In the case of SpCas9, a silhouette score of 0.723 was obtained when using only RNA:DNA hybrid features, while for FnCas9, this score was 0.528 (shown in Figure 4.5E). Silhouette scores were further computed for each native and base-pair mismatch pair. For SpCas9, silhouette scores of 0.671, 0.532, and 0.624 were observed for the native-mm20, native-mm16_19, and native-mm17-20 pairs, respectively. In contrast, FnCas9 exhibited relatively higher values of 0.875, 0.655, and 0.866 for the corresponding pairs. Although, the overall silhouette score of clusters in FnCas9 was lesser than SpCas9, the pair-wise scores of on-target (native) with off-target forms were found to be better in FnCas9. These results indicate that the heteroduplex conformations in on-target form were more distinct from the off-target systems of FnCas9 and possess a lesser degree of separation and distinctiveness among them as compared to those in SpCas9. Furthermore, the pairwise comparisons indicate that the heteroduplex conformation in the native system of FnCas9 is particularly well-distinguished from its off-target systems compared to the corresponding systems in SpCas9. This infers that the heteroduplex conformational changes in FnCas9, unlike SpCas9, are independent of the number and position of PAM-distal mismatches.

4.3.2 State-Dependent Dynamics Unveils Restrained Conformational Changes in FnCas9 Heteroduplexes

The dynamics of the RNA:DNA heteroduplex under mismatches has been extensively studied where the mismatch-induced nucleic acid distortions were reported to be associated with Cas9 activation and off-target tolerance. The significance of understanding the long-term kinetics of the RNA:DNA heteroduplex is vital because it sheds light on changes in conformational dynamics of catalytic domains, influencing Cas9's activation and the off-target tolerance mechanism. In contrast to basic PCA analysis, which offers a limited view of the conformational space of the heteroduplex, long-term kinetics analysis, such as Markov State Modeling (MSM), allows the identification and exploration of discrete conformational states that persist for a significant amount of time in the trajectory (metastable states), their transitions, and their evolution over extended

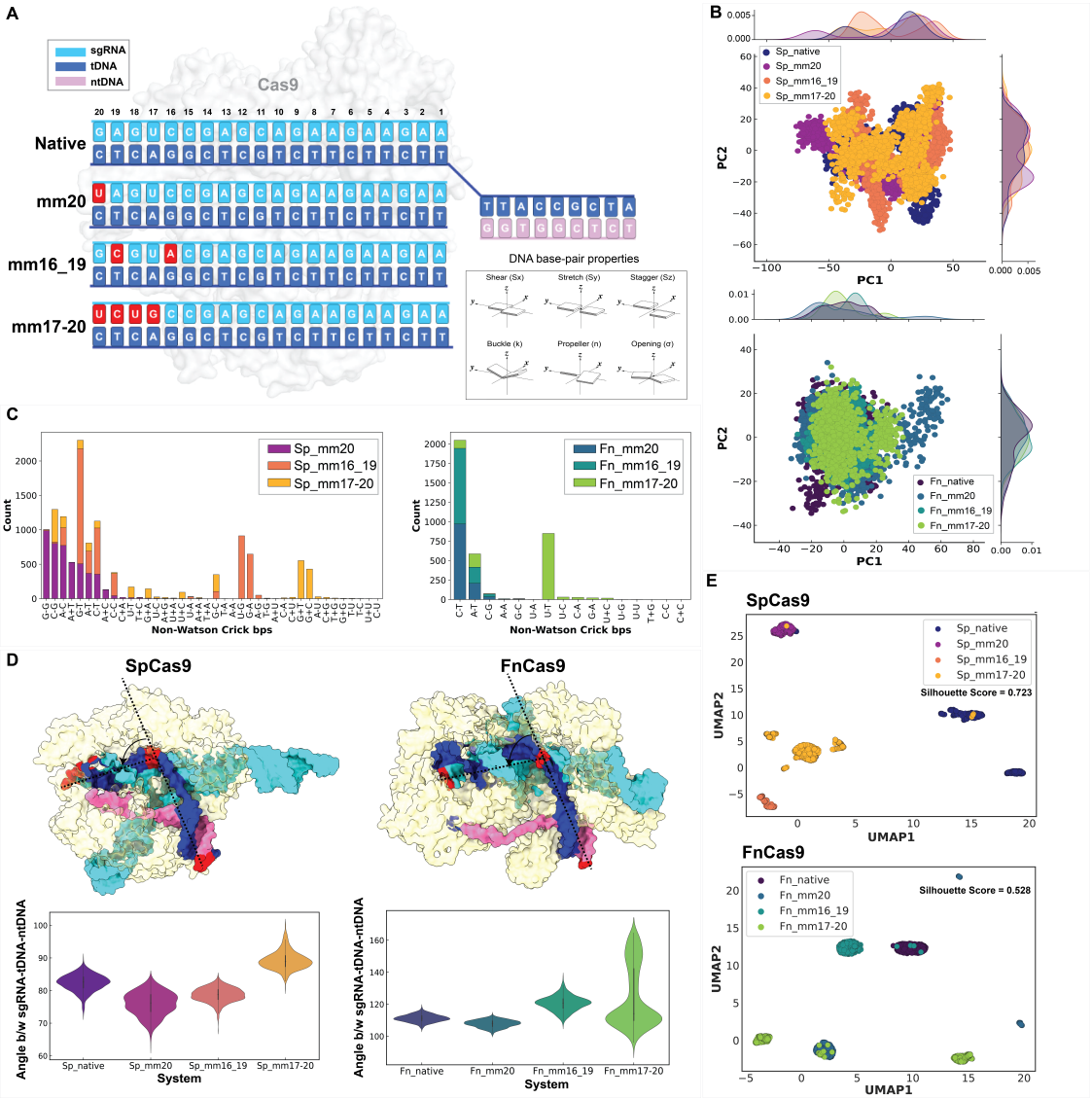


Figure 4.5: A. Diagram of the sgRNA (cyan color) and tDNA (blue color) heteroduplex in CRISPR/Cas9, showing the location of base pair mismatches (mm) at PAM distal sites. The model systems considered for Gaussian accelerated MD (GaMD) simulations including the on-target DNA sequence (on-target) and DNA sequences containing base pair mismatches at PAM distal sites are shown. (i.e., mm20, mm16_19, and mm17–20). Geometrical base-pair descriptors used for explaining the DNA dynamics are shown. B. Principal Component Analysis (PCA) graphs illustrating the conformational distributions of the sgRNA:tDNA heteroduplex in on-target and off-target forms of SpCas9 (top) and FnCas9 (bottom). C. Stacked bar plot showing the distribution of non-canonical base pairs found in off-target systems of SpCas9 (left) and FnCas9 (right). D. The Cas9 ternary complex is shown as isosurface representations that represent the angle between PAM-proximal and PAM-distal duplexes (θ). Violin plot showing the distribution of angle in on-target and off-target systems of SpCas9 and FnCas9. E. Two-dimensional uniform manifold approximation and projection (UMAP) generated using only geometric base pair descriptors for the SpCas9 (top) and FnCas9 (bottom) systems.

intervals of time. The characterization of the metastable states of RNA:DNA hybrid and associated structural changes helps to elucidate the dynamic behavior of the system, offering clues about the

molecular events driving Cas9 activation or influencing off-target tolerance.

To characterize structural heterogeneity, representative structures corresponding to the identified states were extracted and visualized in 2D representations using the RNApdbee tool (Figure 4.6A) Zok *et al.* (2018). SpCas9 exhibited greater structural distortion compared to FnCas9, with more frequent base-pair loss and non-canonical base-pair formation. In FnCas9, all states except State-0 maintained Watson-Crick base-pairing with minimal distortion in the heteroduplex shape. Particularly, in SpCas9, State-2 displayed the least number of Watson-Crick base pairs, with several non-canonical base pairs and a distorted heteroduplex structure as can be seen in Figure 4.6B.

The state population analysis revealed that State-4 and State-3 were the predominant states in SpCas9, constituting 39.7%, and 35.9%, respectively (Figure 4.6B). The proportions of other states in SpCas9 were State-0 (5.7%), and State-1 and State-2 shared equal proportions (4.2%). In FnCas9, State-3 was the most abundant at 61.9%, followed by State-2 (22.9%), State-1 (9.9%), and State-0 (5.3%) (Figure 4.6C). The base-pair complementarity was analyzed by computing the percentage loss of Watson-Crick base pairs in every state identified for the SpCas9 and FnCas9 systems. The State-2 in SpCas9 exhibited a maximum of $\sim$34% base-pair loss and State-3 and State-0 exhibited $\sim$32% base-pair loss (Figure 4.6B). This observation was concurrent with the 2D representation shown in Figure 4.6A, where Sp-S2 had the maximum missing base pairs, followed by Sp-S3 and Sp-S0. FnCas9 states displayed comparatively less loss of Watson-Crick base pairs and fewer non-canonical base pairs (Figure 4.6C). State-0 in FnCas9, the most unstable state, had approximately 10% base-pair loss, contributing only 5.3% to the overall FnCas9 data, indicating less structural disruption of heteroduplex in FnCas9 (Figure 4.6C).

Contributions of on-target and off-target systems to each macrostate in SpCas9 and FnCas9 were determined and visualized in pie charts (Figure 4.6B and 4.6C). In SpCas9, State-2 was mainly composed of frames from the Sp_mm20 system (95.8%) with no contribution from Sp_native. In the other disrupting states of SpCas9, State-3, and State-0, with around 32% base pair loss, Sp_mm17-20 was the main contributor at 33. 8% and 31. 1%, respectively. Fn_mm17-20 contributed significantly to State-0 in FnCas9, representing 94.7% of the state, while State-1 ($\sim$2% base-pair loss) had Fn_mm16_19 as the major contributor. The transition network was also explored by plotting representative structures of the metastable states to comprehend the transition probability between these states. In both the kinetic networks, it can be observed that the unstable states tend to transition to a relatively stable state. The conditional probability of transition was higher for the Sp-S2 to Sp-S3 and Fn-S0 to Fn-S3 states of the SpCas9 and FnCas9 systems, respectively (depicted in Figure 4.6B and Figure 4.6C). The Markov State models aided in highlighting the distinct states identified for FnCas9 and SpCas9 heteroduplexes. This analysis emphasized that the states of FnCas9 remained relatively intact, indicating its restricted conformational changes compared to those of SpCas9 suggesting that FnCas9 heteroduplexes remain structurally unaffected by PAM-distal mismatches, indicating an off-target tolerance independent of heteroduplex shape. It was also observed that the most unstable states in both SpCas9 and FnCas9 had no contribution from native states and higher contributions from the mm20 and mm17-20 states.

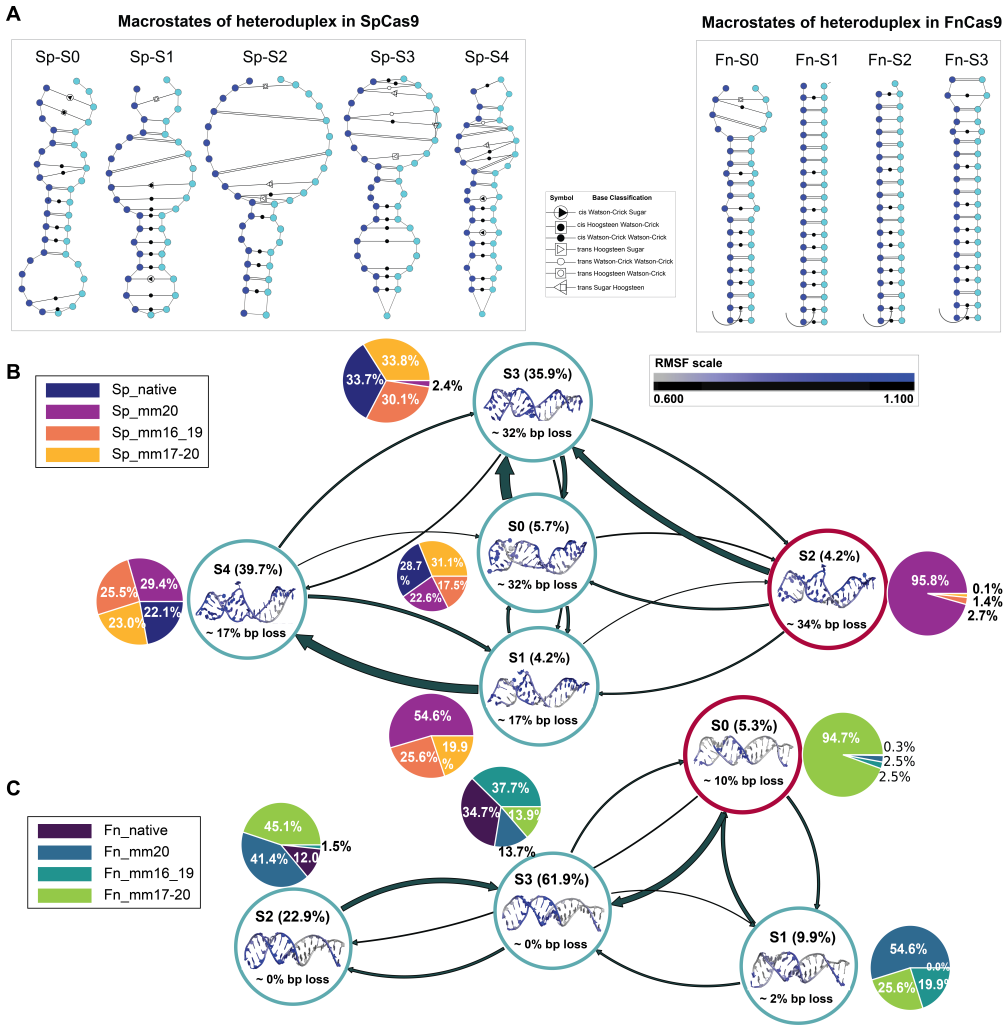


Figure 4.6: A. A 2D representation of the five metastable macrostates found in SpCas9, namely Sp-S0, Sp-S1, Sp-S2, Sp-S3, and Sp-S4; and four metastable macrostates in FnCas9, namely Fn-S0, Fn-S1, Fn-S2, and Fn-S3, generated by the RNApdbee tool Zok *et al.* (2018) using representative structures of each input state. B, C. MSMs illustrating the kinetic network for the SpCas9 and FnCas9 systems. Circles correspond to the macrostates, labeled by the macrostate codes reported in Figure A. The red circle corresponds to the most unstable state with the highest percentage of base-pair loss. The proportion of every macrostate is shown inside the circle. Structural representation of every state is provided inside the circle where the color represents the atomic fluctuations found in that state, ranging from 0.6 (gray) to 1.1 (blue). The gray arrows between macrostates indicate the presence of a probability flux between this pair of states at equilibrium, and the arrow thickness is proportional to the flux. Large gray arrows correspond to kinetically favorable transitions. Pie charts corresponding to every state represent the percentage composition of on-target and off-target forms of SpCas9 and FnCas9 in each state. Each on-target and off-target form of Cas9 is color-coded.

4.3.3 Enhanced mismatch-induced Cas9 domain deviations imply greater sensitivity and specificity

Molecular dynamics simulations uncovered that the presence of four PAM-distal mismatches induced substantial conformational rearrangements of the REC3 domain involving the insertion

of the 692-700 alpha-helix into the heteroduplex, contributing to the extended opening of the RNA:DNA hybrid Ricci *et al.* (2019). Similarly, other simulations demonstrated the correlation between HNH helices and tDNA bases, demonstrating how mismatches at the distal ends of the PAM clearly impact HNH dynamics compared to the positions upstream Mitchell *et al.* (2020). The involvement of RuvC loops in stabilizing mismatch-induced duplex conformations emerged as crucial for Cas9 activationBravo *et al.* (2022*a*). Structural perturbations in the REC3 domain resulting from mismatches at specific positions were also highlighted, disrupting Cas9's conformational checkpointPacesa *et al.* (2022). These findings highlight the impact of mismatch positions on Cas9's domain dynamics, stressing the importance of studying these dynamics in understanding Cas9's mismatch tolerance.

To assess the impact of base-pair mismatches on the conformational dynamics of the SpCas9 and FnCas9 proteins, the root mean square deviation (RMSD), the radius of gyration (ROG), and solvent-accessible surface area (SASA) of the protein backbone atoms were monitored over time. The analysis revealed that the backbone RMSD of both on-target and off-target forms of SpCas9 and FnCas9 stabilized after approximately 300 ns of simulation (Figure 4.1). To elucidate the domain fluctuations observed in the on-target and off-target forms of Cas9, the root mean square fluctuations were calculated by residue. Both Cas9 proteins showed higher fluctuations in the HNH and REC2 domains, particularly in the SpCas9 mm16_19 system and the FnCas9 mm20 and mm17-20 systems (Figure 4.7A). To study the impact of these deviations in the folding dynamics and compactness, the protein's radius of gyration (ROG) was computed (Figure 4.1E-F). Notably, base-pair mismatches at the 16th and 19th positions of SpCas9 led to decreased compactness, while in FnCas9, the mm16_19 and mm17-20 systems exhibited increased compactness than their native form. The average compactness of on-target and off-target forms of both Cas9 proteins remained comparable (Figure 4.7B). FnCas9, being larger in size, exhibited lower compactness than SpCas9, resulting in a larger solvent-exposed surface area evident from the solvent-accesible surface analysis (Figure 4.1G-H)).

To further determine the dominant conformational changes exhibited by the Cas9 domains, PCA analysis was performed for the domains in the recognition (REC) and nuclease (NUC) lobe of Cas9 (Figure 4.7C). In SpCas9, diverse conformational changes were observed in the HNH, RuvC, and REC2 domains across various off-target systems. Differences in Cas9 domain conformational distributions between on- and off-target systems along PC1 and PC2 were assessed by performing the Mann-Whitney U tests. The HNH domain displayed distinct states in SpCas9's off-target systems along PC1, with the mm20 system exhibiting more pronounced deviations. These observations infer that FnCas9 has more flexible HNH and less flexible RuvC domain than SpCas9. Higher domain deviations are found in the HNH domain in the presence of base pair mismatches in both Cas9s unlike the RuvC domain where the deviations occurred irrespective of the base pair mismatches. The REC2 domain in FnCas9 remains almost unaffected despite the base pair mismatches while in SpCas9 different conformations were observed in the on-target and off-target forms. On the contrary, REC3 exhibited statistically significant differences in the conformational distribution between on-target and off-target forms in FnCas9 with higher domain deviations than

SpCas9. Further studies using domain deletion approaches will shed light on the difference in functioning of domains and its affect on the activity and kinetics of both Cas9s.

To assess the effect of off-target binding on the Cas9 domains, the interactions (within 4 Å) established by the RNA:DNA hybrid with the catalytic (HNH and RuvC) and recognition domains (REC1, REC2, and REC3) were determined. In SpCas9 systems, since the REC2 and HNH domains are not in direct contact with the RNA:DNA hybrid, the contacts formed by these domains were comparatively less than the RuvC, REC1 and REC3 domains (800). In FnCas9, a large number of novel contacts were established by REC3 and REC1 domains with RNA:DNA hybrid, suggesting their crucial role in communicating the information to catalytic domains from the RNA:DNA hybrid. To assess structural disparities induced by mismatches in SpCas9 and FnCas9 on- and off-target complexes, protein structural properties like RMSD, ROG, SASA, and domain interactions with the hybrid were used as input to UMAP (Figure 4.7D). The silhouette score was found to be 0.416 for SpCas9 systems and 0.504 for FnCas9 systems, depicting a better structural distinction in the on-target and off-target forms of FnCas9. To understand the structural differences of off-target systems with native forms, silhouette scores were computed for each native-base-pair mismatch pair using only protein features. These scores indicate that base pair mismatches have a greater impact on the FnCas9 protein system than on the SpCas9 system, represented by the greater degree of separation among its clusters, implying higher mismatch sensitivity in FnCas9 system.

4.3.4 Mismatches affect hybrid recognition influencing Cas9's specificity

Examining the heteroduplex dynamics induced by mismatches and assessing their impact on Cas9 domain conformations would aid in offering insights into the structural dynamics underlying Cas9 function, thereby assisting in determining key factors influencing specificity and tolerance to base-pair mismatches. To determine these structural and to elucidate the collective effect of hybrid RNA-DNA and Cas9 protein features on structural variations within CRISPR systems, UMAP visualization techniques were applied to both on-target and off-target forms of SpCas9 and FnCas9 (Figure 4.9A). Interestingly, the clusters formed using these features demonstrated a higher degree of segregation for FnCas9 in comparison to SpCas9 depicted by silhouette scores, with FnCas9 achieving a score of 0.917 and SpCas9 at 0.865. To quantitatively assess the UMAP clusters observed using (a) only RNA:DNA hybrid parameters (28), (b) only protein parameters (14), and (c) both RNA:DNA hybrid, and protein parameters (42), the silhouette score was evaluated and compared. In FnCas9, these scores were found to be 0.525, 0.504, and 0.916 while considering only RNA:DNA hybrid, only proteins, and both hybrid and protein descriptors, respectively, whereas for SpCas9, these values were 0.723, 0.416, and 0.865. The pair-wise silhouette scores computed for every on-target and off-target pair of SpCas9 and FnCas9 systems revealed better on- and off-target discrimination in the case of FnCas9 than SpCas9. The integration of RNA:DNA hybrid and protein descriptors significantly improved inter-cluster separation in both the Cas9s, indicating the significant role of Cas9 domains in distinguishing on-target and off-target systems. In coherence

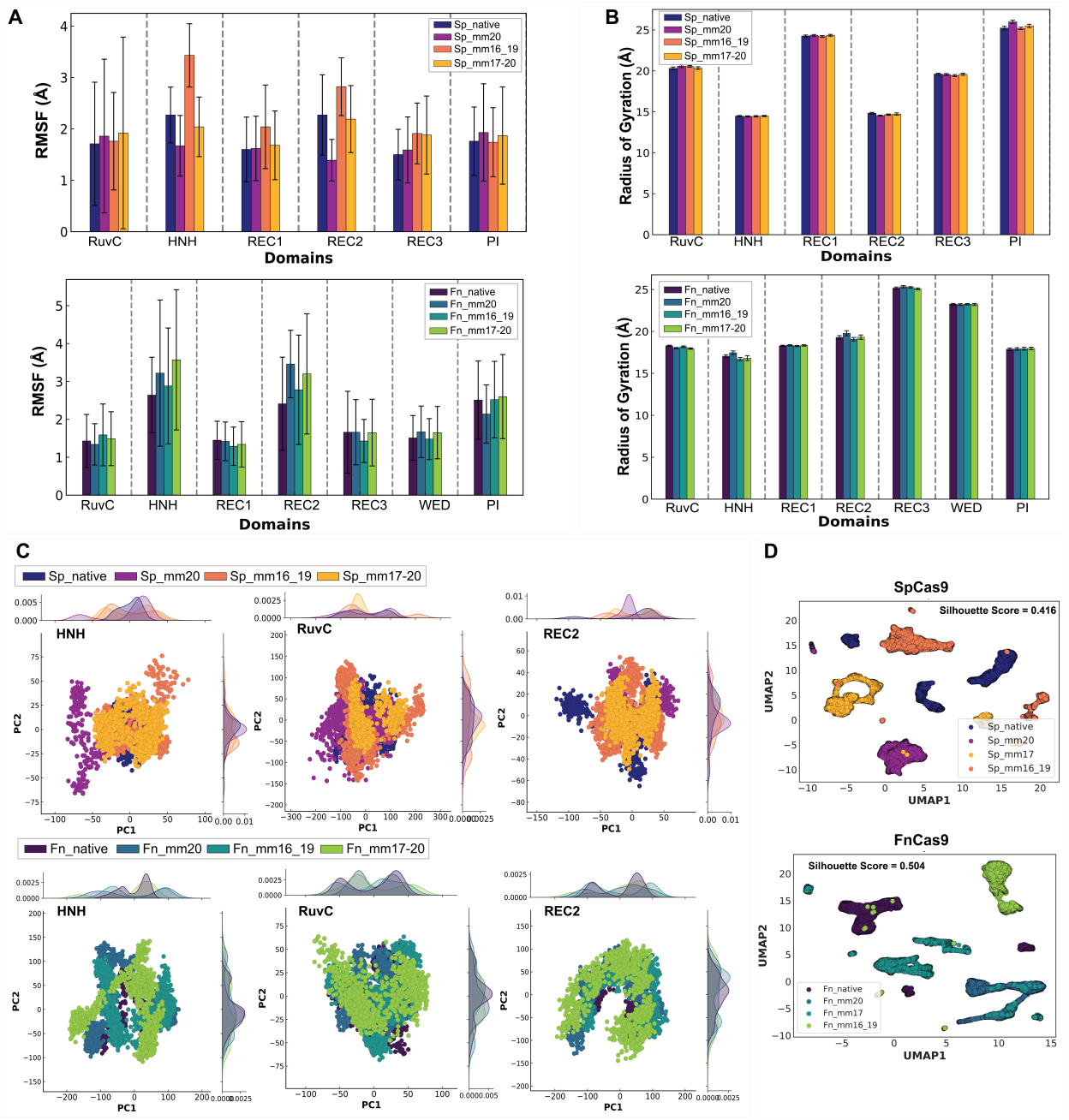


Figure 4.7: A. Average Protein Root Mean Square Fluctuations (RMSF) in angstroms (Å) calculated for the stable part of the trajectory (300-500 ns) representing the RuvC, HNH, REC1, REC2, REC3, and PI domains in both SpCas9 (top) and FnCas9 (bottom) on-target and off-target systems. B. Average Protein Radius of Gyration (Å) within the stable segments of the trajectory (300-500 ns) corresponding to the RuvC, HNH, REC1, REC2, REC3, and PI domains in SpCas9 (top) and FnCas9 (bottom). Data is presented for both on-target and off-target systems. C. Principal Component Analysis (PCA) graphs illustrating the conformational distributions of the HNH, RuvC, and REC2 domains in the on-target and off-target systems of SpCas9 (top) and FnCas9 (bottom). D. Two-dimensional Uniform Manifold Approximation and Projection (UMAP) generated using only protein features for SpCas9 (top) and FnCas9 (bottom) systems

to this observation, incorporation of protein features with DNA:RNA hybrid, led to a substantial improvement in distinguishing the on-target and off-target forms of FnCas9 when compared

against SpCas9. To substantiate domain rearrangement as a significant specificity factor, additional simulation was performed using on-target and off-target forms of more efficient and specific naturally occurring Cas9 ortholog derived from the bacterium Streptococcus thermophilus was simulated. This ortholog exhibits reduced tolerance to base-pair mismatches and recognizes a distinct PAM sequence (NNAGAA/NNGGAA) compared to SpCas9 and FnCas9, which recognize the NGG PAM. Protein and DNA features were extracted and UMAP clustering was performed. The UMAP analysis of StCas9 (Streptococcus Thermophilus) revealed intriguing insights regarding its specificity (Figure 4.9B). Silhouette scores obtained from UMAP clustering indicate significant differences in the contributions of RNA:DNA hybrid and protein parameters. For StCas9, the scores were 0.58 for RNA:DNA hybrid parameters, 0.95 for protein parameters, and 0.75 when both were considered. This suggests that the protein features play a more critical role in distinguishing between on-target and off-target interactions compared to the RNA:DNA hybrid alone. These results align with previous findings for SpCas9 and FnCas9, which demonstrated that protein properties had a greater effect on specificity than the RNA:DNA hybrid. The silhouette scores obtained using only protein parameters for StCas9 (0.95), FnCas9 (0.504) and SpCas9 (0.416) correlated well with the observed order of specificity (StCas9 > FnCas9 > SpCas9) highlighting the significance of domain rearrangement in determining specificity, with higher silhouette scores associated with greater ability to discriminate on-target from off-target interactions. To determine the structural characteristics regulating the metastable states (identified by implementing markov state modeling) of a heteroduplex, machine learning technique was employed where the identified MSM states were used as labels, and the protein and RNA:DNA hybrid structural parameters for each of the states were extracted and used as input features. By characterizing the different states of the RNA:DNA hybrid, it is possible to identify state-specific structural properties of heteroduplex and Cas9 domains, providing insights about features enabling mismatch tolerance in Cas9. This approach isolates key factors, providing mechanistic insights into the RNA:DNA hybrid and protein's individual roles while unraveling their combined impact and highlighting their complex interplay. Given the inherent non-linear relationships anticipated between the extracted structural and dynamic features and the discrete metastable states defined by our Markov State Model (MSM), we prioritized non-linear classification models.

The scores of best performance models for SpCas9 and FnCas9 heteroduplex state classification are shown in Tables 4.1 and 4.2. In the first case, by using only hybrid RNA:DNA features, SpCas9 outperformed FnCas9 (Table 4.1). The XG Boost model achieved the best F1 score of 0.975 for SpCas9 (All Features), while the XG Boost model achieved F1-score 0.880 for FnCas9 by using RFE feature selection method. While, FnCas9 performed better using protein structural parameters (F1 score (XG Boost)-0.963) compared to SpCas9 (F1 score (Grad. Boost)-0.910). In the third case, both RNA:DNA hybrid and protein structural features were jointly considered, resulting in improved model performance for both Cas9s. The best performance for SpCas9 (F1 score-0.986) was achieved using the BorutaSHAP feature selection method with the gradient-boosting model, while FnCas9 (F1 score-0.965) achieved its best performance using the extra-tree feature selection method for the XG Boost Machine learning model (Table 4.1 and Table 4.2). These outcomes suggest that on considering a combination of RNA:DNA hybrid and Cas9 do-

Performance using only RNA:DNA hybrid Features					
Feature Selection Method	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
All features	XG Boost	0.975	0.975	0.975	0.975
RFE	XG Boost	0.971	0.972	0.971	0.971
Extra-Tree Classifier	XG Boost	0.973	0.974	0.973	0.974
BorutaSHAP	XG Boost	0.974	0.974	0.974	0.974
Feature Selection Method	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
All features	Gradient Boosting	0.911	0.910	0.911	0.910
RFE	Gradient Boosting	0.911	0.910	0.911	0.910
Extra-Tree Classifier	XG Boost	0.836	0.831	0.836	0.831
BorutaSHAP	Gradient Boosting	0.911	0.909	0.986	0.910
Feature Selection Method	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
All features	Gradient Boosting	0.985	0.985	0.985	0.985
RFE	Gradient Boosting	0.984	0.984	0.984	0.984
Extra-Tree Classifier	XG Boost	0.984	0.984	0.984	0.984
BorutaSHAP	Gradient Boosting	0.986	0.986	0.986	0.986

Table 4.1: Machine learning model performance of SpCas9 System using Protein, DNA and Protein+DNA features using RFE, Extra-Tree Classifier and BorutaSHAP feature selection methods

main features, SpCas9 demonstrates better ability in classifying heteroduplex metastable states compared to FnCas9 systems. Additionally, the marginal difference between the F1-scores pre- and post-feature selection for both Cas9s suggests that the choice of feature selection method had a limited impact on the overall model's performance. These outcomes suggest that on considering a combination of RNA:DNA hybrid and Cas9 domain features, SpCas9 demonstrates better ability in classifying heteroduplex metastable states compared to FnCas9 systems. Additionally, the marginal difference between the F1-scores pre- and post-feature selection for both Cas9s suggests that the choice of feature selection method had a limited impact on the overall model's performance.

To investigate important feature types impacting heteroduplex state classification, a comparative analysis of common features identified by RFE, BorutaSHAP, and Extra-Tree Classifier feature selection methods was conducted for both Cas9 systems, as depicted in Figure 4.9C. The analysis revealed 15 common features in SpCas9 and 17 common features in FnCas9 across all three feature selection techniques. Among the common features in SpCas9, six were protein structural parameters, while in FnCas9, 12 out of the 17 common features were associated with Cas9 protein structural parameters. These findings emphasize the significance of protein parameters in FnCas9 heteroduplex state classification. To identify the primary features that contribute to RNA:DNA hybrid classification, a feature ranking analysis was performed using BorutaSHAP and the extra tree classification methods shown in Figure 4.9D-E . Common features selected by BorutaSHAP for both Cas9s, along with their mean importance scores, are presented in Figure 4.9D. Importance scores given by extra tree classifier method for the 42 features of both Cas9s are shown in Figure

Performance using only RNA:DNA hybrid Features					
Feature Selection Method	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
All features	Gradient Boosting	0.901	0.880	0.901	0.880
RFE	XG Boost	0.904	0.886	0.904	0.880
Extra-Tree Classifier	Gradient Boosting	0.894	0.878	0.894	0.847
BorutaSHAP	XG Boost	0.902	0.881	0.964	0.878
Feature Selection Method	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
All features	XG Boost	0.964	0.963	0.964	0.963
RFE	Gradient Boosting	0.911	0.910	0.911	0.910
Extra-Tree Classifier	XG Boost	0.836	0.831	0.836	0.831
BorutaSHAP	XG Boost	0.964	0.963	0.964	0.963
Feature Selection Method	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
All features	XG Boost	0.966	0.964	0.966	0.965
RFE	Support-vector machine	0.965	0.964	0.965	0.965
Extra-Tree Classifier	XG Boost	0.966	0.965	0.966	0.965
BorutaSHAP	Support-vector machine	0.958	0.956	0.958	0.957

Table 4.2: Machine learning model performance of FnCas9 System using Protein, DNA and Protein+DNA features using RFE, Extra-Tree Classifier, and BorutaSHAP feature selection methods

4.9E. The difference in the importance scores of features in SpCas9 and FnCas9 was evaluated using both the Extra-Tree classifier and BorutaSHAP feature selection methods (Figure 4.8). A threshold, set at 1.5 times the standard deviation from the mean of the differences in importance scores, was applied to identify key features. Features exceeding this threshold were highlighted with yellow and blue boxes in Figures 4.9D and 4.9E.

In the BorutaSHAP ranking, non-native contacts between the RuvC domain and the RNA:DNA hybrid were significant for SpCas9 (blue box in Figure 4.9D), while FnCas9 showed significant features including native contacts by HNH with the hybrid, non-native contacts by REC3 with the hybrid, and protein backbone RMSD (yellow box). In the Extra-Tree classifier ranking, significant features for SpCas9 included native contacts of RuvC with the hybrid and non-native contacts by HNH and RuvC (blue box in Figure 4.9E). For FnCas9, crucial features were non-native contacts formed by REC3 with the hybrid and protein backbone RMSD (yellow box). Both methods indicated that non-native contacts formed by the RuvC domain with the RNA:DNA hybrid were important for heteroduplex state classification in SpCas9, while protein backbone RMSD and non-native contacts between the REC3 domain and RNA:DNA hybrid were significant for FnCas9.

4.3.5 Shorter allosteric pathway ensures efficient information transfer enhancing specificity

To validate the structural significance of the identified crucial features from the feature ranking, an assessment of the binding affinity between the heteroduplex and the REC (REC1, REC2, and

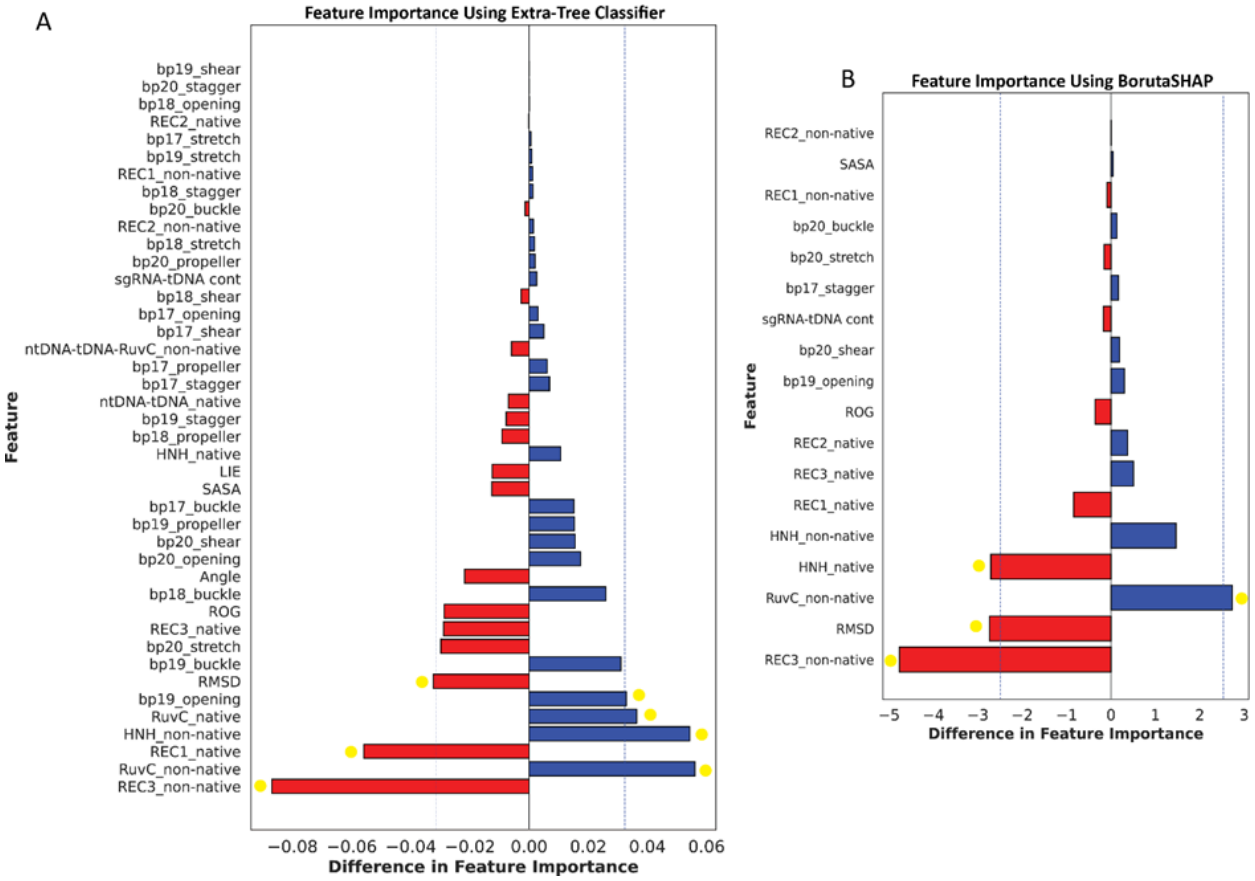


Figure 4.8: Difference plot showing difference in (protein+hybrid) feature importance obtained using Extra-Tree classifier and BorutaSHAP method. Threshold based on 1.5 times the standard deviation from the mean of the difference in importance scores. Threshold for Extra-Tree method: +/- 0.032 and BorutaSHAP method: +/- 2.485. Red bars represent features that are more important for FnCas9, while blue bars represent features more important for SpCas9. Key features based on threshold are highlighted using yellow dot.

REC3 domains) along with the NUC (HNH and RuvC) lobes was conducted using MM/GBSA analysis. This analysis included all metastable states derived from the SpCas9 and FnCas9 systems. The evaluation of binding affinity revealed that in SpCas9, the nuclease lobe exhibited a considerably tighter binding affinity to the heteroduplex. Conversely, the recognition lobe demonstrated a more pronounced binding affinity in the FnCas9 systems, highlighting a structural divergence between the two Cas9 systems (as depicted in Figure 4.10A). Comparative analysis of binding energy distributions for the SpCas9 and FnCas9 systems revealed a statistically significant difference (p-value < 0.05), demonstrating the distinctive nature of their interactions (Figure 4.10A). Per-residue energy decomposition analysis was performed to find residues that contribute to the binding of heteroduplex states to the REC and NUC lobes of the SpCas9 and FnCas9 systems. The assessment of crucial residues for FnCas9 and SpCas9 involved plotting the residue-wise energy distribution. Identification of key residues was achieved by applying a threshold at the 5th percentile of the distribution resulted in threshold energy of -6.09 kcal/mol and -5.13 kcal/mol for the recognition lobe and -5.40 kcal/mol and -4.82 kcal/mol for nuclease lobe in SpCas9 and FnCas9, respectively. Based on these cut-off energy values, the key residues in SpCas9

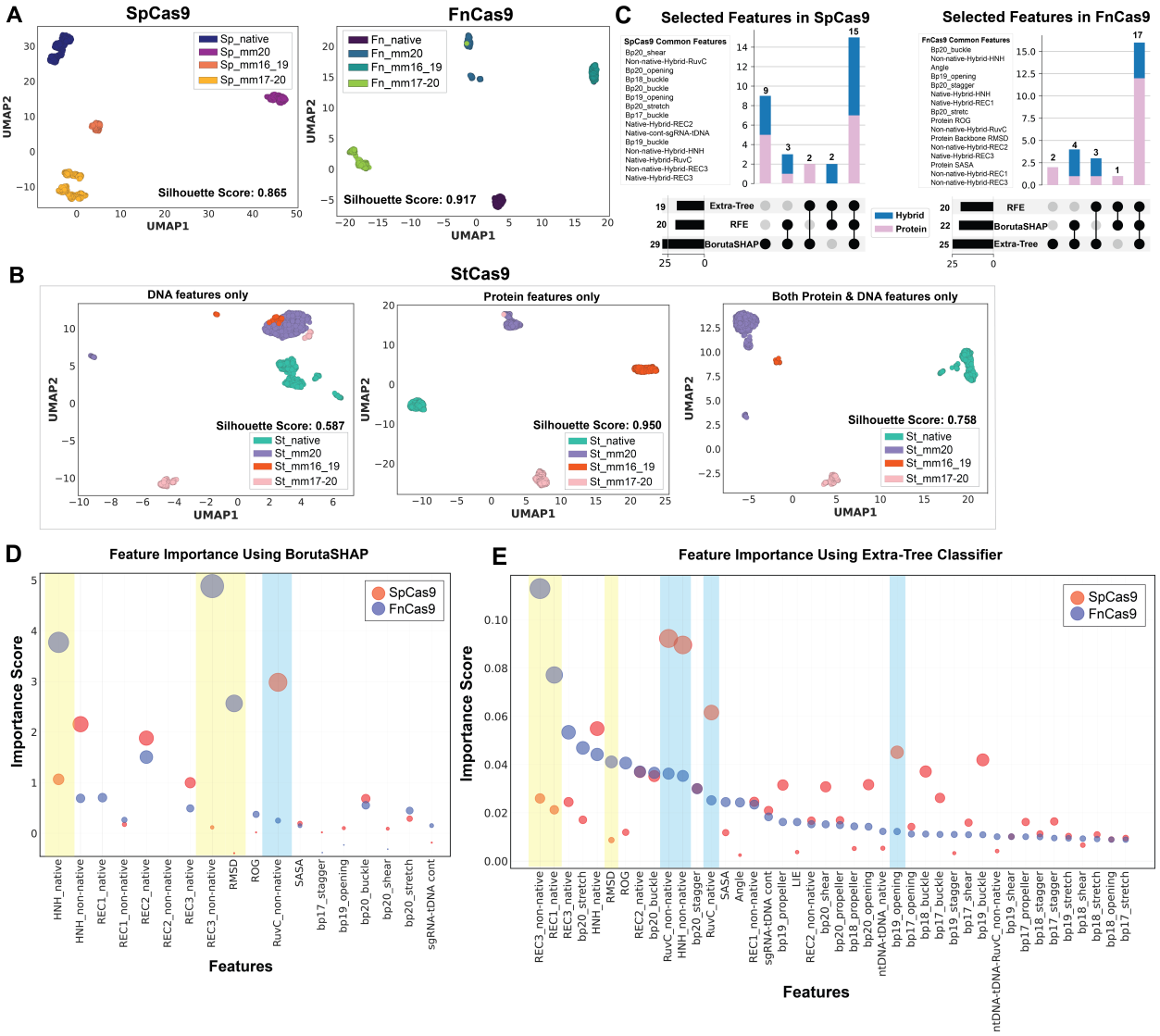


Figure 4.9: A. Two-dimensional projection of Uniform Manifold Approximation and Projection (UMAP) components, for SpCas9 (left) and FnCas9 (right) systems using RNA:DNA hybrid and Protein structural parameters. B. UMAP clusters obtained for StCas9 (*Streptococcus Thermophilus*) using only DNA, only protein and both DNA & protein features. C. Upset plot depicting the intersecting features identified through Recursive Feature Elimination (RFE), Extra-Tree Classifier, and BorutaSHAP feature selection methods in SpCas9 and FnCas9. D. Bubble plot showing feature importance scores of SpCas9 (red) and FnCas9 (blue) using the BorutaSHAP algorithm for intersecting selected features. E. Bubble plot showing feature importance scores of SpCas9 (red) and FnCas9 (blue) using extra-tree classifier method for all features.

included: REC - Tyr450, Lys526, Arg661, Phe405, Asn588, Hie167, Met694; NUC - Arg919, Arg925, Arg765, Lys1014. The key residues in FnCas9 were REC - Phe104, Arg622, Arg721, Arg849, Tyr794; NUC - Val858, shown in Figure 4.10B. The residues of the REC2 domain were found to be least involved in the binding of the heteroduplex, especially in the FnCas9 protein. The non-involvement of REC2 residues in the heteroduplex binding hinted at its allosteric role in Cas9's functioning. Several reports have emphasized REC2's role as an allosteric transducer in transmitting information from the REC3 recognition site to the HNH catalytic site in Cas9

proteins. To elucidate the precise mechanisms underlying information transfer between Cas9 domains, a comparative analysis of allosteric pathways was conducted in Checkpoint and pre-cleavage states of SpCas9 in contrast with FnCas9. In a Cas9 variant, HypaCas9 residues in REC3 domain are mutated (R692A/M694A/Q695A/H698A or HypaCas9) Chen *et al.* (2017*a*) resulting in more specific Cas9. These mutations altered allosteric signaling, which hindered the REC-II re-orientation and obstructed the HNH activation in the presence of PAM distal mismatches. Leveraging these observations, N773/R692 in the REC3 domain was considered as source residues, and H903/H840 in the HNH domain was considered as sink residue for determining the shortest pathway in checkpoint/pre-cleavage states. Similarly, in FnCas9, a REC3 residue proximal to the PAM-distal region was designated as the source, with catalytic residue N995 in HNH serving as the sink. Intriguingly, the comparative analysis revealed distinctive pathway lengths between SpCas9 and FnCas9. In SpCas9, the shortest pathways, consisting of 13 nodes/residues in the checkpoint state (Checkpoint state path: 773, 602, 670, 638, 342, 334, 318, 315, 305, 948, 955, 916, 903) and pre-cleavage state (Pre-cleavage state path: 692, 521, 516, 592, 597, 651, 270, 264, 248, 832, 859, 822, 839), were identified. Conversely, in the FnCas9 ternary complex, the shortest pathway comprised only 8 nodes (FnCas9 path: 840, 836, 1041, 1048, 984, 988, 949, 995), as illustrated in Figure 4.10C. In both the checkpoint and precleavage states of SpCas9, the shortest pathway linking the recognition and catalytic sites involved residues within the REC2 domain, indicating their role in facilitating information transmission. On the contrary, the comparative analysis revealed a contrasting pattern in FnCas9, where the REC2 domain did not contribute to information transmission. This distinction led to a notably shorter, faster, and more efficient information transfer mechanism observed in FnCas9, potentially enhancing its specificity.

4.4 Conclusion

Elucidating the dynamics and structural alterations within CRISPR/Cas9 systems is pivotal for unraveling their functional disparities and off-target implications. FnCas9 is reported to have high target specificity with minimal off-target binding, attributed to its low DNA binding affinity in the presence of target sgRNA mismatches. It produces staggered cleavage, higher homology-directed repair rates, and lower nonspecific genome editing compared to SpCas9. This study aimed to decipher the factors behind its higher specificity by analyzing heteroduplex dynamics influenced by PAM-distal base-pair mismatches and their impact on Cas9 domain dynamics and RNA:DNA interactions. These dynamical changes impact signal transmission mechanisms within domains, offering insights into factors modulating information transfer and protein specificity. PCA and MSM, revealed restricted conformational changes in FnCas9 heteroduplexes compared to SpCas9 systems. Interestingly, FnCas9 exhibited greater conformational deviations in protein domains despite base-pair mismatches implying that its cleavage mechanism is predominantly dependent on Cas9 domain rearrangement rather than the heteroduplex shape. UMAP clustering showed that FnCas9 better distinguishes on-target and off-target forms compared to SpCas9 when combining hybrid and protein features. This suggests that FnCas9's discrimination is more dependent on hy-

brid contacts with Cas9 domains rather than hybrid shape. To further validate this observation, we took StCas9, a protein previously shown to have strong structural signatures for the editing properties. The UMAP analysis of StCas9 further reinforces the significance of domain rearrangements in determining Cas9 specificity, indicating that protein features play a more critical role in distinguishing on-target from off-target interactions compared to the RNA:DNA hybrid alone. The machine-learning classification of heteroduplex states revealed a slightly higher classification accuracy in SpCas9, suggesting more distinguishable conformational states. It is also highlighted that classification in FnCas9 depends on intricate interactions between its domains and the hybrid. While in SpCas9, it relies more heavily on the intrinsic structural features of the hybrid alone, indicating a less domain-interactive approach. Feature ranking revealed that REC1/REC3-hybrid contacts were important for FnCas9 hybrid classification whereas HNH/RuvC-hybrid contacts were crucial in the case of SpCas9.

The MM/GBSA results further reinforced these findings, revealing higher binding affinity of the RNA:DNA hybrid in SpCas9 to the nuclease (NUC) lobe, contrasting with FnCas9 where stronger binding occurred to the recognition (REC) lobe. Consequently, these observations suggest that base-pair mismatches exert differential impacts, influencing NUC-hybrid binding in SpCas9 and REC-hybrid binding in FnCas9. Allosteric signal transmission in SpCas9 involved the REC3, REC2 and HNH domains, whereas in FnCas9, direct transmission from REC3 to HNH occurred, bypassing REC2 due to its distance from the heteroduplex. FnCas9 exhibited a shorter pathway for information transmission, potentially enhancing specificity compared to SpCas9. Based on these observations, the potential response mechanism of FnCas9 to mismatches appears to involve the role of REC3 in recognizing and inhibiting HNH activation, thus trapping the system in a checkpoint state. On the contrary, in SpCas9, REC3 appears to relay information to REC2, impeding HNH activation in the presence of mismatches. Apart from this, there could be several factors behind the higher specificity of FnCas9, such as – staggered cleavage in FnCas9, the presence of structural Zinc ion, and different orientation of the PAM-interacting domain than SpCas9.

This study highlights the intricate interplay between mismatch positions and Cas9's conformational dynamics, providing crucial insights into the dependencies on specific Cas9 domains and the direct information transmission which opens up a promising avenue for optimizing Cas9 design involving targeted mutagenesis or deletion experiments (REC2 domain). Whilst FnCas9 and SpCas9 are the main focus of this work, it's essential to consider other engineered high-fidelity Cas9 variants that contain amino-acid substitutions to enhance specificity by allosterically modulating Cas9 function and locking HNH in an inactive state. Variants like eSpCas9, SpCas9-HF1, HypaCas9, and evoCas9 with mutations near the PAM-distal end affect information transmission from REC3 to HNH via REC2. In contrast, xCas9 3.7, Sniper-Cas9, SpCas9^{2Pro}, and HifiCas9 have mutations in REC3, RuvC, HNH, and PI domains, altering DNA targeting specificity differently. It remains elusive regarding how these variants allosterically modulate target DNA cleavage and mismatch tolerance. The basis behind designing these variants signifies the importance of domain-rearrangement and the non-specific interaction formed between RNP complex-tDNA or sgRNA-tDNA in regulating the specificity. Future research should aim to compare the important

structural factors uncovered in this study with high-fidelity Cas9 variants and validate it experimentally to understand their mechanism of cleavage and factors behind higher specificity. By integrating insights from various Cas9 variants, we can further optimize Cas9 design and improve the precision and efficiency of genome editing technologies.

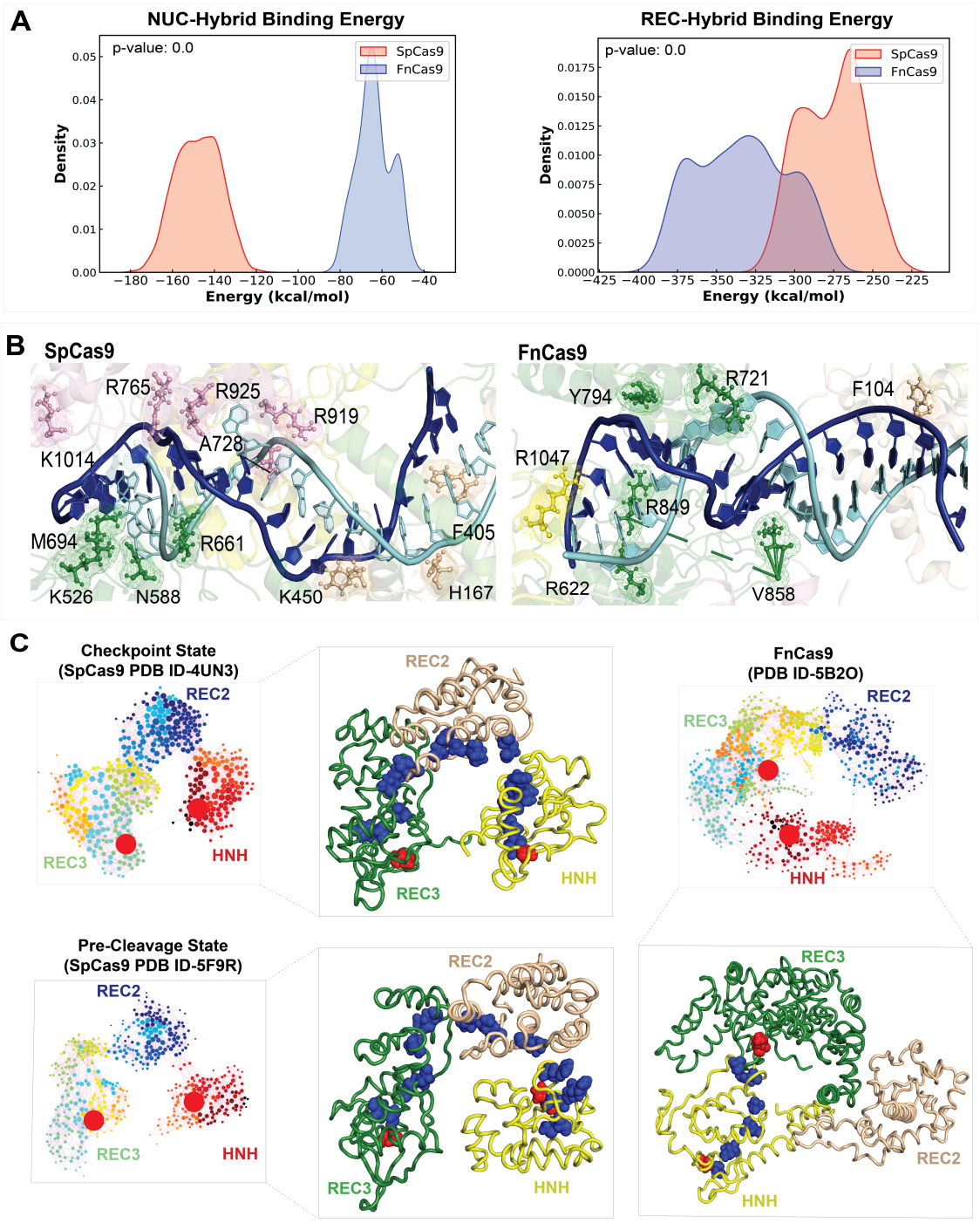


Figure 4.10: A. Distribution of binding affinity between nuclease (NUC)-sgRNA:tDNA heteroduplex and recognition (REC)-sgRNA:tDNA heteroduplex found in SpCas9 and FnCas9 system. B. Key residues involved in the binding of heteroduplex with REc and NUC lobes in SpCas9 and FnCas9 proteins are depicted in stick representation. Color representation of sgRNA-cyan; tDNA-blue; HNH residues-yellow, RuvC residues-pink; REC3 residues-green, REC2-residues-beige color. C. Residue interaction network found between REC3-REC2-HNH domains shown in SpCas9 - checkpoint and pre-cleavage states and FnCas9 ternary complex. The recognition and catalytic sites of REC3 and HNH domains are shown using red spheres. These sites were used as source and sink residues to find the shortest path. REC3, REC2, and HNH domains are shown in dark green, beige and yellow colors in tube representation. The residues involved in the shortest path are shown in the close-up view as blue spheres.

CHAPTER 5

Structural Insights into Mismatch Recognition by Engineered FnCas9 Variants

5.1 Introduction

FnCas9, derived from *Franscisella novicida*, a well-known CRISPR-Cas9 variant has a number of major advantages over the widely used SpCas9 that is derived from *Streptococcus pyogenes.* Although both enzymes recognize the identical minimal 5'-NGG-3' protospacer adjacent motif (PAM) sequence, FnCas9 exhibits significantly enhanced specificity contingent on the single-guide RNA (sgRNA) sequence when assessed with DNA substrates. Whereas high-fidelity variants of SpCas9 engineered to mitigate off-target effects have been developed, their genome editing efficiency frequently remains markedly inferior to that of the wild-type enzyme. To address this issue, researchers have pursued diverse alternative microbial Cas systems for genome editing applications; however, none have consistently exhibited better efficiency compared to SpCas9. Previous investigations by Debojyoti, and group Acharya *et al.* (2019*b*) have indicated that FnCas9 possesses extremely high inherent specificity, easily disassociating from off-target sites during in vitro experiments. Conversely, SpCas9 and its high-fidelity variants remain attached to off-targets without cleaving them, which might result in unintended off-target effects. Even in base editing scenarios, where cutting is not required, the limitations of the targeting window of Cas9 base editors have led to the development of Cas9 variants with modified PAM specificities to reach desired targets. They noted that FnCas9 exhibits a lower tolerance to mismatches, ceasing cleavage when more than a single base pair mismatch is present. Additionally, they found that FnCas9 is less efficient than SpCas9 and its larger size may hinder its delivery. To address these limitations, the same group rationally engineered FnCas9 by modifying its WED-PI domain and phosphate-lock loop (PLL), creating enhanced (en) FnCas9 variants (en1, en15 and en31) (Figure 5.1) without compromising its intrinsic specificity Acharya *et al.* (2024). They also demonstrate robust cellular editing, improved HDR-mediated knock-in rates, and single nucleobase specificity in human cells. Furthermore, the compatibility of the en31 variant with extended (x-) or super-extended (sx-)-gRNAs allows tuning of the adenine base editing window for ABEs without requiring cumbersome PAM engineering. This attribute broadens the range of pathogenic SNPs that can be targeted by en31-ABE. In this work, Debojyoti and group have also worked on REC2 deleted version of the protein(Δ S112-A297). In my previous study (Chapter 4), it was observed that FnCas9 follows a shorter route for information transmission between catalytic and cleavage site as REC2 was not involved in this transmission. The domain arrangement of FnCas9 is slightly different from SpCas9 and they took advantage of this fact and evaluated the the cleavage activity of truncated en1FnCas9 by deleting its REC2 domain. The removal of the REC2 domain was

a strategic decision as it reduced the overall size of the protein, bringing it closer in size to SpCas9, which can enhance its delivery and expression in cellular systems, potentially improving its efficacy. Remarkably, the truncated en1 variant (en1-ΔREC2) retained DNA cleavage efficiency comparable to the full-length en1 variant, indicating that the REC2 domain may not be essential for cleavage activity in this specific variant. In this study, en1 (E1369R), en15(E1603H) and en31 (E1369R, E1449H, G1243T) were having the highest cleavage activity among other variants (Figure 5.1). The en31 variant is a triple mutant (G1243T/E1369R/E1449H) that initiates DNA unwinding through specific hydrogen bonding and electrostatic interactions with the DNA backbone. Notably, en31 demonstrates broad activity at altered PAM sites and excels in distinguishing single mismatches. It is particularly compatible with super-extended gRNAs, allowing for broader targeting capabilities. According to this study, en1FnCas9 and en31FnCas9 appear to be promising candidates for enhancing gene-editing capabilities. To gain a deeper understanding of the structural dynamics and specificity of these engineered FnCas9 variants, we conducted molecular dynamics (MD) simulations of en1, en15, and en31 in both their native and base-pair mismatch forms. By analyzing the interactions between the Cas9 variants and their target DNA, as well as their behavior in the presence of mismatches, we aim to elucidate the stability and flexibility of the protein-DNA complex and how it relates to the specificity of the Cas9 proteins. Additionally, structural comparison with truncated engineered variants (removed REC2) will further enhance the understanding about the structural rearrangement that takes after REC2 removal and how the it alters the domain and heteroduplex dynamics. Studying theses conformational and structural changes in the presence of PAM-distal and PAM proximal mismatch would further help in advancing our knowledge how the the position of mismatch affects the heterduplex dynamics and binding and can this be correlated with specificity of the Cas9. This study would also help in determining if the REC2 domain plays a critical role in stabilizing the protein-DNA complex or if its absence leads to improved kinetics and specificity. This comparative analysis can enhance our understanding of the structural dynamics of these proteins and inform the design of more effective genome editing tools.

5.2 Materials and Methods

5.2.1 Structural Modeling and Conventional MD simulations

Atomistic molecular dynamics simulations have been performed using the X-ray structure of *Francisella novicida* Cas9 (PDB ID: 5B2O) Hirano *et al.* (2016) in complex with RNA and DNA, solved at 1.7 Å resolution. In the case of FnCas9, the non-target strand (ntDNA) was extended by ten nucleotides, utilizing the ntDNA strand of SpCas9 as a template. The nucleotides on the sgRNA strand of FnCas9 were mutated to target the human EMX1 gene sequence (hEMX1) (5′-GAGUCCGAGCAGAAGAAGAA-3′). One base-pair was added at the 21st position distal to PAM, to make 21-nt long sgRNA:tDNA duplex. Using this structure as template, two variants of FnCas9 – en1FnCas9 (E1369R) and en31FnCas9 (E1369R, E1449H, G1243T) were also mod-

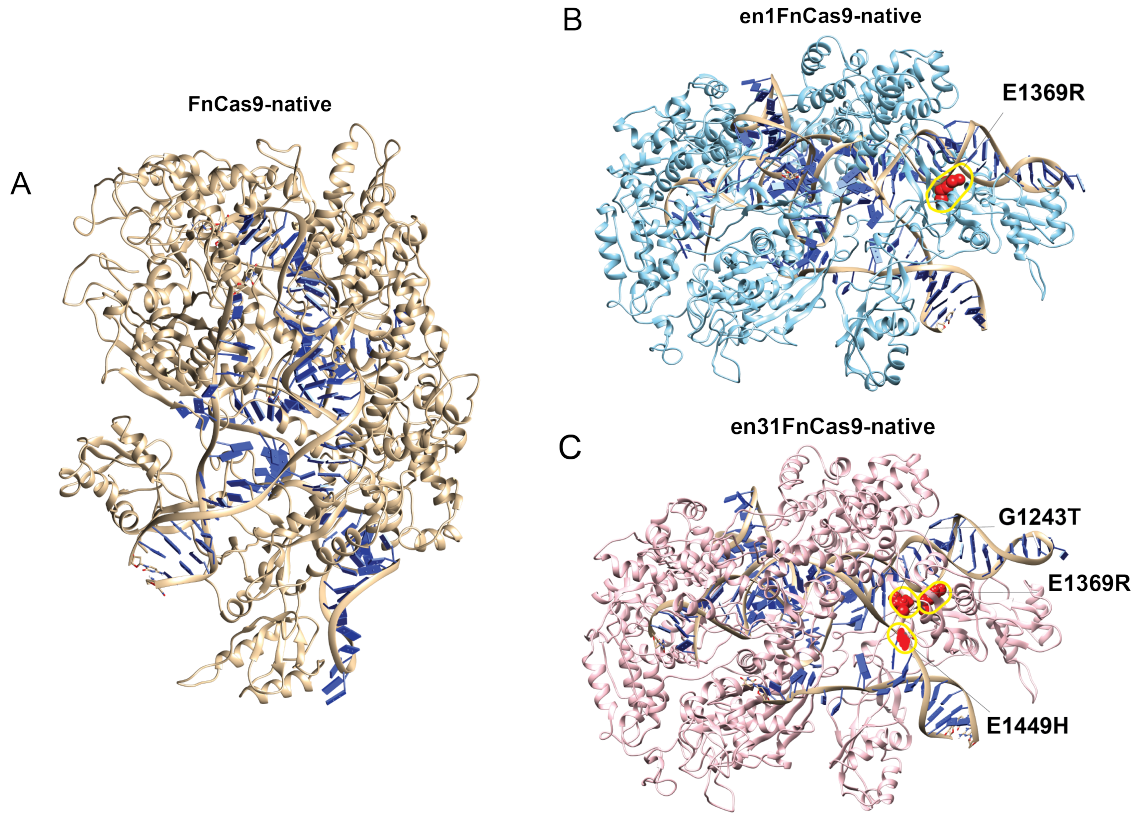


Figure 5.1: Crystal structure of the F. novicida and it's engineered variants bound to a guide RNA, the target DNA, and nontarget DNA strand. A. FnCas9 (native), B. en1FnCas9 (E1369R), C. en31FnCas9 (G1243T, E1369R, E1449H)

eled using SwissModel Schwede *et al.* (2003). Off-target systems were also modeled for all three systems (FnCas9-native, en1FnCas9, and en31FnCas9) by introducing base pair mismatches at PAM-distal end of the sgRNA:tDNA heteroduplex G:U (at position 20), A:C (at position 19), G:U (at position 18), U:G (at position 17), C:A (at position 16), and C:A (at position 2) which was proximal to the PAM. Following our observation made in previous Chapter and recent work by Debojyoti and group Acharya *et al.* (2024), truncated en1 and en31FnCas9 with REC2 removed were also modeled and simulated by following the similar protocol. Also, in these REC2 deleted models, a proximal mismatch (at position 2) was also introduced to compare its affect on heteroduplex and protein's conformations. All the systems which were simulated are listed in the Figure5.2. All constructed model systems were minimized in Chimera Pettersen *et al.* (2004) and used as starting structures for performing simulations. These were immersed in explicit water with the addition of Na+ ions to maintain charge neutrality, creating orthorhombic periodic simulation cells, containing ∼300,000 atoms for the FnCas9 system. Simulations were performed in explicit water with Na+ ions for charge neutrality, using the Amberff14SB Maier *et al.* (2015) force field with specific corrections for DNA and RNA, and the Aqvist model for Mg2+ ions. The systems were solvated in explicit water with Na+ ions for charge balance, creating orthorhombic periodic cells with approximately 300,000 atoms for FnCas9. Equilibration was performed using conventional MD with the Amberff14SBMaier *et al.* (2015) force field, including specific corrections for DNA and RNA, and the Aqvist Aqvist (1990) model for Mg2+ ions. Energy minimization relaxed

water molecules and counter-ions, while the system was gradually heated from 0 to 100 K in the NVT ensemble with 50 kcal/mol restraints for 100 ps, followed by an increase to 200 K in the NPT ensemble with 25 kcal/mol restraints for ∼100 ps. Releasing all restraints, the temperature was raised to 300 K in a single 500 ps NPT simulation. The production run was conducted for 50 ns in the NPT ensemble using the GPU version of AMBER22 HMADA *et al.* (2022). The equilibrated systems were used for GaMD simulations.

5.2.2 Gaussian Accelerated MD simulations

The system potential energy surface (PES) is smoothed using the improved sampling technique known as "accelerated MD" (aMD), which reduces energy barriers and speeds up transitions between low-energy states by introducing a non-negative boost potential Hamelberg *et al.* (2004). In GaMD, the boost potential applied follows a Gaussian distribution, which is a more robust and novel way of accelerating simulations Miao *et al.* (2015). The boost potential has been applied in a dual-boost scheme, in which two acceleration potentials are applied simultaneously to the system: (i) the torsional terms only and (ii) across the entire potential. A time step of 2 fs was used. The maximum, minimum, average, and standard deviation values of the system potential (V_{max}, V_{min}, V_{avg}, and σ_V) were obtained from ∼50 ns cMD with no boost potential. Each GaMD simulation proceeded with a ∼25 ns run, in which the boost potential was updated every 1.6 ns, thus reaching equilibrium values. Finally, ∼ 250 ns of GaMD simulations were performed in the NPT ensemble. For each system, ∼250 ns (GaMD production) + 25 ns (GaMD equilibration) + 50 ns (pre-equilibration conventional MD) was carried out. Since GaMD has been applied on 8 model systems, including on and off-target DNAs, a total of ∼3.9 μs (i.e., ∼250 ns of production + 75 ns of equilibration × 18 systems) of simulations has been produced. The MD trajectory snapshot was saved after every 1 ns. Only the stable part of the trajectory was used for simulation analysis (200-250 ns). These simulations were performed with the GPU version of AMBER22.

5.2.3 MD Simulation Trajectory Analysis

Geometrical properties, including the RMSD of protein backbone atoms (Å) and RNA:DNA hybrid, domain-wise radius of gyration and residue-wise RMSF were also computed for both the Cas9s in their on-target and off-target forms. Trajectory maps (TrajMap) was used to visualize protein's backbone movements during the simulation as a heatmap to visualize protein's behavior over time, compare Fn-native simulation with its variants using difference maps Kožić and Bertoša (2024). The sgRNA:tDNA structural dynamics were analyzed during the GaMD production runs using the 3DNA code Lu and Olson (2003). Loss of base pairing and the count of non-canonical base pairs formed within the RNA:DNA hybrid in mismatch systems for both Cas9 proteins were also assessed, utilizing the 3DNA code. To investigate the dominant motions and conformational distributions of RNA:DNA heteroduplex and Cas9 domains, the covariance matrix was built and diagonalized using the Pytraj module Roe and Cheatham III (2013). The analysis was performed

using the cartesian coordinates for the sgRNA-tDNA heteroduplex, RuvC, HNH, REC1, REC2, and REC3 domains. The analysis used all frames from all equilibrated simulations in on-target and off-target forms of both Cas9s. To determine key residues which are involved in making tight contacts (within 4 Å) with the heteroduplex in >80% frames in the stable part of the trajectory. These residue-nucleotide pairs are represented as a network using NetworkX Hagberg *et al.* (2008) package.

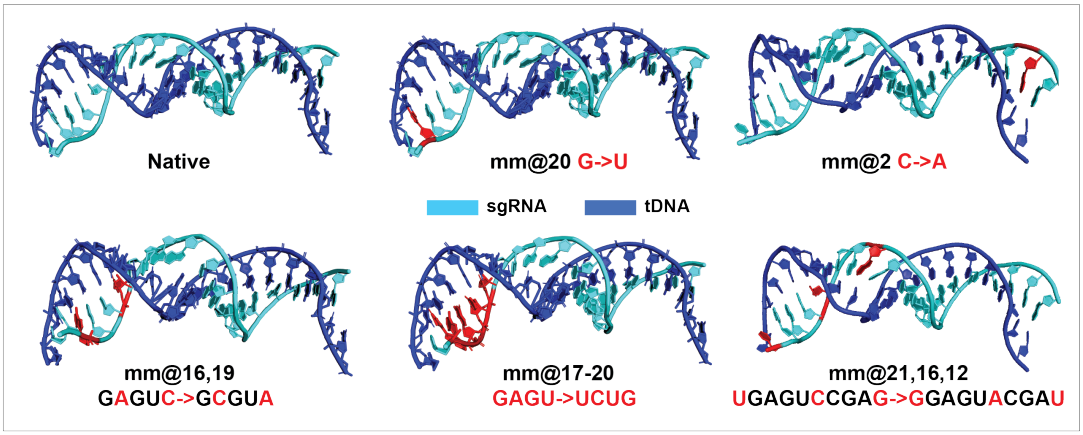
5.2.4 MM/GBSA Calculations

The Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method Genheden and Ryde (2015) is a popular technique for determining binding free energies between proteins and ligands. gmx_MMPBSA Valdés-Tresanco *et al.* (2021) was used to calculate the binding free energy of the nuclease (NUC) and recognition (REC) lobes with RNA:DNA hybrid. Both the lobes were used as receptors, and heteroduplex was used as the ligand. Molecular mechanics potential energy (electrostatic and van der Waals energies), the free energy of solvation (polar and nonpolar solvation energies), and the energetic contribution of each residue and nucleotide to total binding energy were obtained. The gas-phase energies calculated using the molecular mechanics force field are a sum of van der Waals (VDWAALS) and electrostatic (EEL) energies, whereas the solvation energy includes a polar (EGB) and a non-polar (ESURF) component.

5.2.5 Clustering Analysis

Clustering analysis was performed using cpptraj, a module of the AmberTools suite, to investigate the conformational diversity of the RNA:DNA heteroduplex. A total of 250 frames from the stable portion of the molecular dynamics trajectory were selected for analysis. The K-means clustering algorithm was employed to categorize these frames into three distinct clusters. The representative frame from the dominant cluster was extracted for further examination. To visualize the heteroduplex region, specifically the 21-nucleotide (nt) segment, a two-dimensional representation was generated using RNApdbee, a tool designed for RNA secondary structure visualization. This representation allowed for a clear depiction of the structural features and interactions within the heteroduplex, facilitating a better understanding of its dynamics and stability in both on-target and off-target contexts.

A



B

Fn-native (21nts)	en1Fn-native (21nts)	en31Fn-native (21nts)
Native	Native	Native
Mm20	Mm20	Mm20
Mm16,19	Mm16,19	Mm16,19
Mm17-20	Mm17-20	Mm17-20
	Mm2	Mm2
	ΔREC2	ΔREC2
	ΔREC2-mm2	ΔREC2-mm2
1300ns	**2275ns**	**2275ns**

Figure 5.2: A. Diagram of the sgRNA (cyan color) and tDNA (blue color) heteroduplex in CRISPR/Cas9, showing the location of base pair mismatches (mm) at PAM distal sites. Model systems considered for Gaussian accelerated MD (GaMD) simulations including the on-target DNA sequence (on-target) and DNA sequences containing base pair mismatches at PAM distal sites are shown. (i.e., mm20, mm16 19, mm17–20, and mm2).
B. Model systems used for simulations are listed along with the total simulation time as each system was simulated for 325ns (*18 systems).

5.3 Results and Discussions

5.3.1 Conformational landscape and domain flexibility in engineered Fn-Cas9 variants

The specificity of engineered FnCas9 variants stems from the DNA interaction properties of Fn-Cas9 itself, rather than the engineered residues at the PAM-duplex region Acharya *et al.* (2024). Among the 31 engineered variants of enFnCas9, en1 (E1369R), en15 (E1603H) and en31 (E1369R, E1449H, G1243T) exhibited the highest cleavage activity. The same study compared the activity and specificity of en1FnCas9 and en31FnCas9 with other PAM-flexible and high-fidelity Cas9 variants, including SpRY39 and Superfi-Cas9. The results demonstrate that en31FnCas9 exhibits

superior specificity and efficiency with higher DNA cleavage rates compared to these other variants. To decipher factors behind its higher efficiency and specificity, all-atom GaMD simulations were performed for both native and engineered variants of FnCas9 (en1 and en31) in the presence of base-pair mismatches. To capture the time-dependent fluctuations within FnCas9 variants with respect to its native form, trajectory maps for both native and FnCas9 variants were computed, and their difference with the native form was compared (Figure 5.3A-B). In this plot, the red regions indicate high fluctuation in their native form, while the blue regions indicate high fluctuation in their variant forms. In en1FnCas9, a region belonging to the residue range 300-420 (REC1) was highly fluctuating; however, these fluctuations disappeared in the last few frames. In en31FnCas9, residues 300-360 and 400-440 exhibited high fluctuation throughout the simulations and are associated with the REC1 domain. Domain-wise RMSF analysis revealed that HNH domain was highly fluctuating in en31FnCas9 while REC2 domain was highly fluctuating in en1FnCas9. PI domain has shown higher fluctuations in both the engineered variants of FnCas9 (Figure 5.3C).

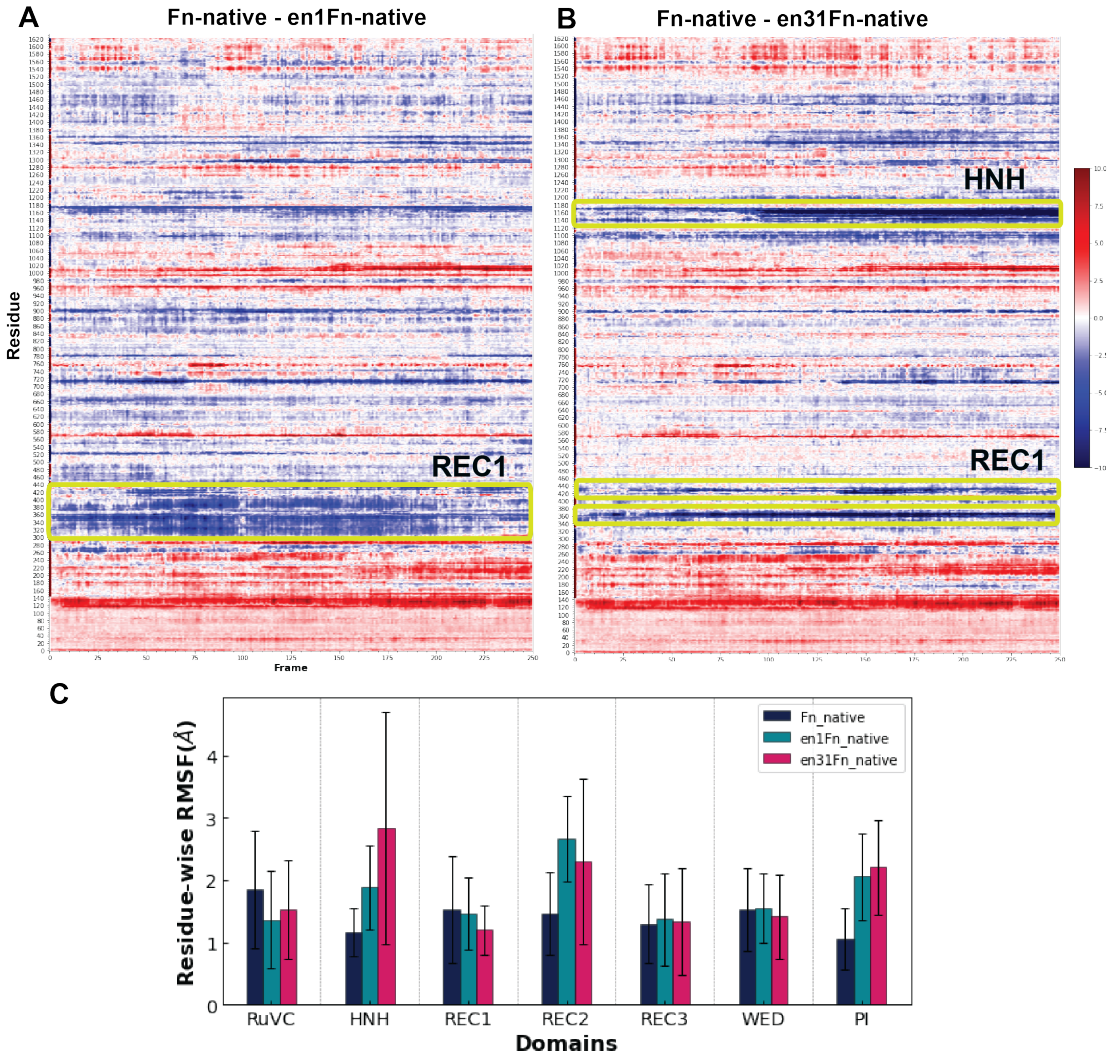


Figure 5.3: A. Trajectory map difference between FnCas9 native and en1FnCas9. REC1 region showing higher shifts in en1FnCas9 (blue) B. Trajectory map difference between FnCas9 native and en31FnCas9. HNH and REC1 region showing higher shifts in en31FnCas9 (blue) C. Residue-wise RMSF line plot for FnCas9 (native-violet), en1FnCas9 (teal), and en31FnCas9 (pink).

Furthermore, domain-wise principal component analysis, and residue-wise RMSF analysis was used to compare the domain-conformational dynamics, flexibility, and compactness of FnCas9 and its variants in their full-length and truncated form (ΔREC2) (Figure 5.4). The residue-wise RMSF analysis was performed to compare highly fluctuating regions in the variants compared to native and how these fluctuations were affected when REC2 was removed (Figure 5.4A). The analysis based on root mean square fluctuation (RMSF) plots and principal component analysis (PCA) scatter plots reveals significant differences in the structural dynamics of various Cas9 enzyme variants. The RMSF plots demonstrate that wild-type FnCas9 exhibits fluctuations across the sequence, with peaks indicating regions of high flexibility. In contrast, the engineered variants en1FnCas9 and en31FnCas9 show distinct fluctuation patterns, suggesting altered dynamic behaviors. Interestingly, the deletion of the REC2 domain in en1FnCas9_delREC2 and en31FnCas9_delREC2 leads to reduced fluctuations, particularly in regions where REC2 is absent, implying structural stabilization. The PCA scatter plots further corroborate these findings (Figure 5.4B). In the full-length forms, the RNA:DNA hybrid of the en1FnCas9 enzyme was observed to exhibit a higher degree of deviations compared to the en31FnCas9 hybrid, which displayed more limited conformational changes. In the HNH domain, the en1FnCas9 variant demonstrated restricted conformational changes, while the en31FnCas9 hybrid showed more pronounced variations, aligning with the findings from the RMSF and shift analyses. Additionally, the RuvC domain experienced more significant conformational changes in variant forms, particularly in en1FnCas9. The REC1 domain revealed a greater conformational distribution in the en1FnCas9 hybrid than in other forms. Interestingly, both the REC2 and REC3 domains exhibited similar conformational changes across all three Cas9 variants, providing a comprehensive understanding of the conformational dynamics within these domains. However, these observations differed when truncated version of en1FnCas9 and en31FnCas9 were taken into account where each variant occupied different region in the principal component space. The deletion of REC2 causes a significant shift in the conformational landscape, particularly in the HNH and REC3 domains. These shifts are evident across various structural components, including the hybrid, REC1, HNH, RuvC, and REC3 domains.

5.3.2 DNA-Protein Interaction Analysis of en1FnCas9 and en31FnCas9 Variants

For both en1FnCas9 (upper plot) and en31FnCas9 (lower plot), the contact profiles between DNA and protein residues are largely preserved between the full-length and REC2-deleted variants (Figure5.5). The network of interactions spans across key functional domains, with dense contacts observed primarily in the REC1, REC3, and WED domains, while interactions with the HNH, and PI domains are also apparent. Notably, the removal of the REC2 domain does not significantly alter the number or distribution of protein-DNA contacts suggesting that by REC2 domain's removal, the protein-DNA interaction profile remains remarkably similar in both the en1FnCas9 and en31FnCas9 variants. This observation suggests that REC2 does not directly contribute to DNA binding, and that its deletion does not impair the protein's ability to form critical contacts with

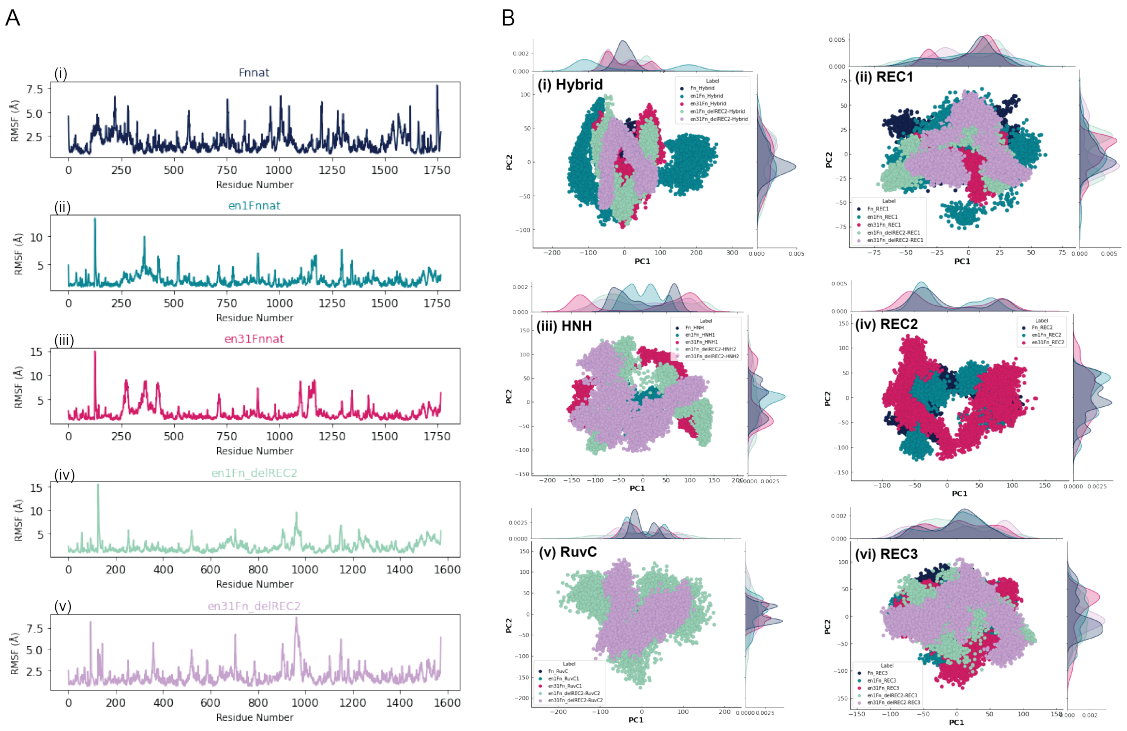


Figure 5.4: A. Residue-wise RMSF for different Cas9 variants (i) FnCas9, (ii) en1FnCas9, (iii) en31FnCas9, (iv) en1ΔREC2, and (v) en31ΔREC2. B. Principal Component Analysis (PCA) graphs illustrating the domain-wise conformational distributions of the (i) Hybrid, (ii) REC1, (iii) HNH, (iv) REC2, (v) RuvC and (vi) REC3 in the native (violet), en1FnCas9 (teal), en31FnCas9 (pink), en1ΔREC2 (seagreen), and en31ΔREC2 (purple).

the DNA substrate. The robust interaction network maintained through REC1, REC3, and WED domains appears to compensate for the absence of REC2, ensuring the structural and functional integrity of the protein-DNA interface. These results are consistent with prior dynamic analyses (PCA and RMSF) showing that while REC2 deletion induces changes in the conformational landscape of the protein, it does not significantly disrupt key functional interactions with DNA. This finding highlights the potential for truncating non-essential regions of the protein, such as REC2, without compromising DNA binding, which could be advantageous for the design of more compact and efficient Cas9 variants.

5.3.3 Cas9-hybrid binding dynamics and residue contributions

To understand how these FnCas9 variants differ in protein-heteroduplex and sgRNA-tDNA binding and what residues and domains have a role to play in its binding, MMGBSA calculations were performed for the stable part of the trajectory (Table 5.1). En31FnCas9 exhibited the highest binding affinity between sgRNA and tDNA (-177.78 kCal/mol). However, in comparing the binding affinity of the Cas9 protein and hybrid native FnCas9 exhibited the highest protein-heteroduplex binding affinity, followed by en31FnCas9 and en1FnCas9 (Table 5.1). Domain-wise binding energies were also calculated to determine the role of each domain in hybrid binding. REC3 domain

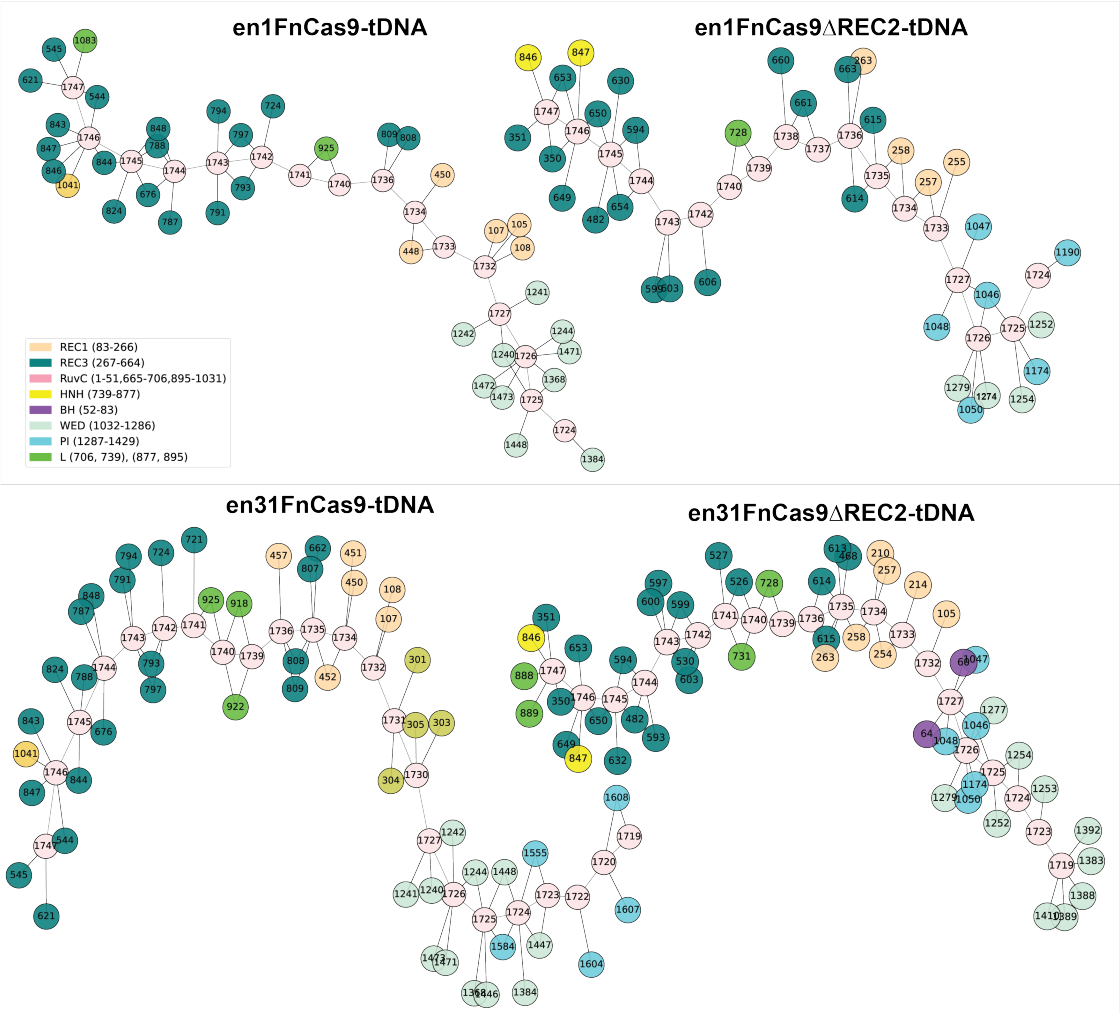


Figure 5.5: Contact plot visualizing the interaction between DNA nucleotides (depicted in pink) and protein residues from different domains of the en1FnCas9 and en31FnCas9 variants, both in their full-length and REC2-truncated forms. The protein residues are color-coded based on their domains, such as REC1 (teal), REC3 (light teal), RuvC (yellow), HNH (purple), BH (blue), WED (green), PI (orange), and the I region (lime).

emerged as the primary contributor to binding across all three systems, especially in en31FnCas9, where it had the highest binding affinity. Surprisingly, the HNH and RuvC domains contributed the least to protein-heteroduplex binding in en31FnCas9. REC3 is the largest domain among all and covers a larger portion of the nucleic acid binding cleft, so its higher content makes sense.

Now, to decipher the key residues involved in Cas9 hybrid binding in all three systems, per-residue energy decomposition and contact analysis per residue were performed (Figure 5.6) In FnCas9 native form and en1FnCas9 major contributors in hybrid binding belonged to L1 and L2 loops (green color in Figure 5.4). In contrast, for en31FnCas9, the residues from the REC3 domain played a greater role in hybrid binding compared to the L1-L2 loops. In the native form of FnCas9, Arg1083 and Asn1084 were identified as pivotal contributors. In the en1FnCas9 variant, Arg1047 and Arg849 emerged as key residues, with significant contributions from the L2 loop. Conversely, for the en31FnCas9 variant, Arg622, Arg721, and Arg849 were determined to be critical residues, all of which are located within the REC3 domain. This underscores the pronounced role of the

System	RNA-DNA	Cas9-Hybrid	REC1-Hybrid	REC2-Hybrid	REC3-Hybrid	HNH-Hybrid	RuvC-Hybrid
Fn-native	-117.65	-575.74	-117.67	-29.08	-193.77	-42.51	-36.93
en1Fn-native	-115.21	-485.95	-76.97	-17.43	-192.91	-46.39	-22.18
en31Fn-native	-177.78	-494.30	-81.65	-21.17	-233.21	-27.71	-17.22

Table 5.1: Binding affinity calculated between sgRNA-tDNA and Cas9-hybrid (domain-wise for HNH, RuvC, REC1, REC2, and REC3) for stable part of trajectory (200-250 ns) using MMGBSA method.

REC3 domain in en31FnCas9 and suggests a distinct substrate recognition mechanism.

These observations are also in agreement with contact analysis, where more residues from REC3 domains contacted the hybrid in en31FnCas9, and fewer contacts were formed by residues from the nuclease lobe (Figure 5.6). In the native FnCas9, the REC1 and REC3 domains established an equivalent number of contacts (18) with the hybrid, followed by the L1-L2 loop (11). Conversely, in the en1FnCas9 variant, the contacts mediated by the REC1, HNH, and RuvC domains were abolished, while the REC3 domain exhibited an increased number of hybrid contacts (27). The RNA:DNA heteroduplex in en31FnCas9 showcased the maximum number of contacts with the REC3 domain (34), followed by the REC1 domain. Other domains in en31FnCas9 exhibited a reduced contribution to heteroduplex binding, as corroborated by MM-GBSA analysis. These results suggest that the higher fluctuations in HNH domain of en31FnCas9 led to its decreased binding affinity with the hybrid and REC3 residues have a role to play in the higher sensitivity of en31FnCas9.

5.3.4 Enhanced off-target discrimination in engineered variants

In the previous sections, we discussed the behavior of the FnCas9 variant and its native forms, highlighting the distinctive features of the en31FnCas9 that may contribute to its higher efficiency or specificity. However, it is important to further investigate how these systems behave in the presence of base-pair mismatches. In a study performed by Jinek's group reported that non-canonical base-pairing enable mismatch tolerance Pacesa *et al.* (2022). To elucidate the role of non-canonical base-pairing in off-target tolerance, the enumeration and classification of non-canonical base pairs were performed for both native and engineered variants of FnCas9,using stable part of the trajectory (Figure 5.7). In non-canonical base-pair types, pyrimidine-pyrimidine base pairs (C-T / U-T) known for their weaker stacking and reduced stability were more common in all off-target system of native and engineered FnCas9 shown in Figure 5.7A-C. The count of these non-canonical base-pair types was much higher in the engineered variants than in the native form suggesting greater distortion in the duplex structure (Figure 5.7A). The off-target system of en31FnCas9 exhibited more diverse non-canonical base-pairing (Figure 5.7C) as compared to off-target systems of FnCas9 (Figure 5.7A) and (en1FnCas9 (Figure 5.7B), especially in four-mismatch system (mm17_20).

The presence of PAM-distal base-pair mismatches resulted in the loss of Watson-Crick base pairs, thereby disrupting the structure of the RNA:DNA heteroduplex (Figure 5.7D-F). On com-

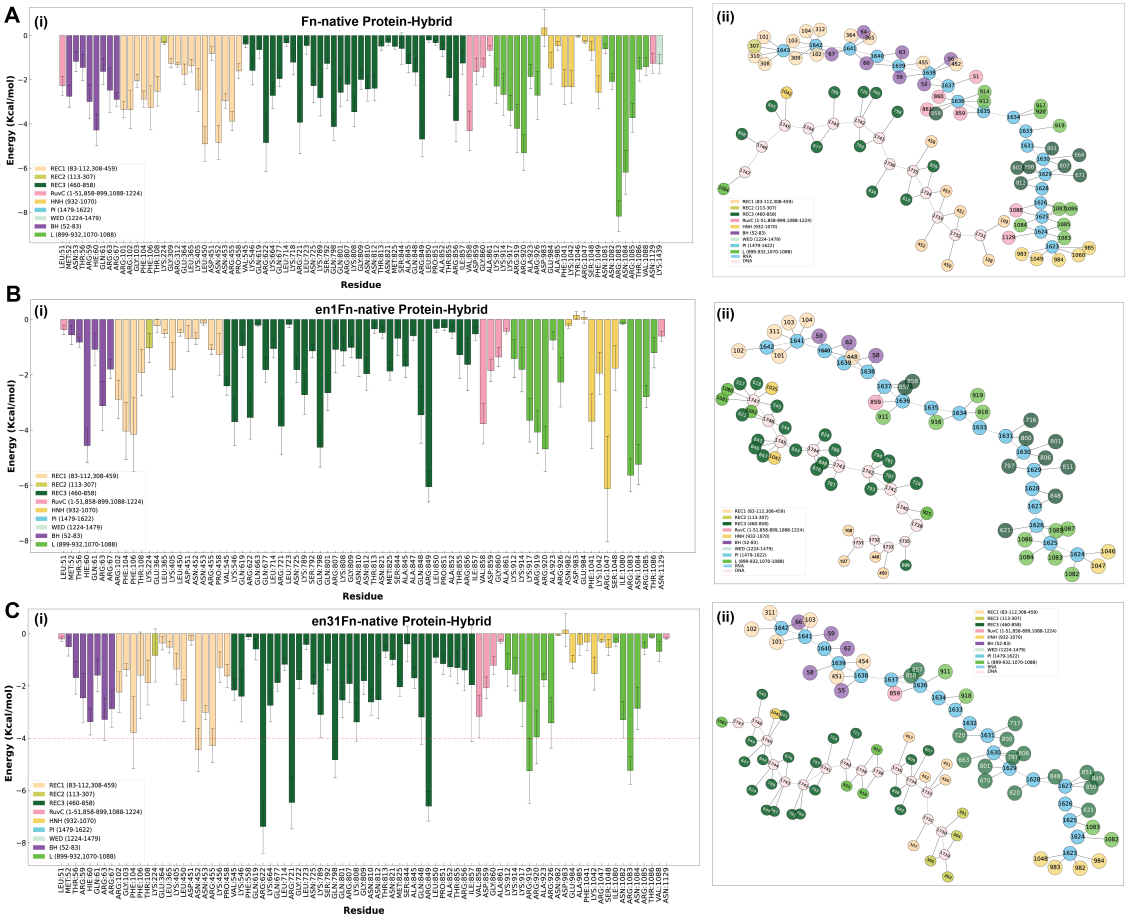


Figure 5.6: Energetic contribution of residues involved in Cas9 and hybrid in A (i)FnCas9(native), B (i) en1FnCas9, and C(i) en31FnCas9. Contact network between Protein-sgRNA, and Protein-tDNA. Contacting residues (within 4Åin more than 80% frames) from different domains are coloured differently A(ii) Contact network in FnCas9 (native), B(ii) Contact network in en1FnCas9, C(iii) Contact network in en1FnCas9.

paring the distribution of base-pair losses across different mismatch conditions, it was observed that the Fn-native (Figure 5.7D) and en1FnCas9 (Figure 5.7E) variants exhibit a narrow distribution with peaks at lower base-pair losses (0-2), indicating minimal impact from mismatches. In contrast, the en31FnCas9 variant (Figure 5.7F) displays a significantly broader distribution with higher base-pair losses extending up to 5, highlighting its increased sensitivity to mismatches. These findings suggest that en31FnCas9 exhibits greater sensitivity to mismatches compared to Fn-native and en1FnCas9, as indicated by more diverse non-canonical base-pairing and higher loss of base-pairs in off-target systems. Additionally, the removal of REC2 disrupted heteroduplex base pairing in both Cas9 variants, with en31FnCas9 showing more base-pair loss (upto 7 base-pair loss) (Figure 5.7F) and non-Watson-Crick pairing (Figure 5.7C) compared to en1FnCas9 (upto 4 base-pair loss). As previously reported for SpCas9, proximal mismatches exhibit more detrimental effect than distal mismatches, compromising heteroduplex stability and hindering R-loop formation and reducing cleavage efficiency. Similar trends were observed for FnCas9, where greater base-pair loss occurred with proximal mismatches than the PAM-distal mismatches. Interestingly, while more non-Watson-Crick base pairing was noted in the presence of PAM-distal mismatches

compared to proximal ones, the loss of base pairs was greater with proximal mismatches. This suggests that although non-Watson-Crick pairs may help maintain duplex integrity and stability, significant base-pair loss can distort the heteroduplex structure and potentially impair Cas9 functionality.

Clustering analysis was performed to extract the dominant conformation of the heteroduplex in the trajectory. The representative frame of the dominant cluster was extracted and used to visualise the 2D representation of the heteoruduplex in on-target and off-target systems of en1FnCas9 and en31FnCas9 in their full-length and REC2 deleted versions too (Figure 5.7G). The 2D representations of heteroduplex provide a clear visual comparison of the structural effects induced by mismatches in various FnCas9 variants. PAM-distal mismatches in the engineered variants (Figure 5.7G (i)) lead to a noticeable widening of the duplex strands, though the presence of non-canonical base pairs, such as Hoogsteen interactions or trans-sugar base pairings, helps maintain some structural integrity. The en31FnCas9 heteroduplex has shown slighlty more structural distortion than the en1FnCas9 heteroduplex.

Contrastingly, mismatches in the PAM-proximal region cause more significant structural disruptions (Figure 5.7G (ii)). The 2D diagrams show that proximal mismatches in both engineered variants result in a greater loss of base pairing, introducing loops and bulges that destabilize the heteroduplex. Particularly, in en1FnCas9, the widening of the heteroduplex was observed in both the ends showing severe structural distortions. This indicates that proximal mismatches pose a greater challenge to the stability of the heteroduplex. Since these mismatches occur near the PAM sequence, where initial target recognition occurs, they have a more substantial impact on the overall stability of the heteroduplex and, in turn, reduce the efficiency of target recognition and cleavage by the FnCas9 complex.

When comparing the ΔREC2 and ΔREC2mm2 systems of both en1FnCas9 and en31FnCas9, noticeable differences emerge. In the case of ΔREC2, which involves the deletion of the REC2 domain, both engineered variants retain more regular base pairing with fewer disruptions and loops, indicating that the removal of REC2 has minimal impact on the base-pairing integrity within the RNA-DNA heteroduplex (Figure 5.7G (iii)). Interestingly, when comparing the proximal mismatches in the full-length and REC2-deleted versions of both engineered variants, the structural instability caused by the proximal mismatches in the full-length variant was mitigated in the REC2-deleted version. The REC2 deletion resulted in less base-pair loss and reduced structural distortion, suggesting a compensatory effect of REC2 removal in alleviating the destabilization typically induced by proximal mismatches. (Figure 5.7G (iv))

5.4 Conclusions

The engineered FnCas9 variant en31 demonstrates better specificity and efficiency compared to native FnCas9 and other high-fidelity Cas9 variants such as SpRY39 and Superfi-Cas9. The present study and structural analysis reveal that en31FnCas9's enhanced sensitivity to mismatches

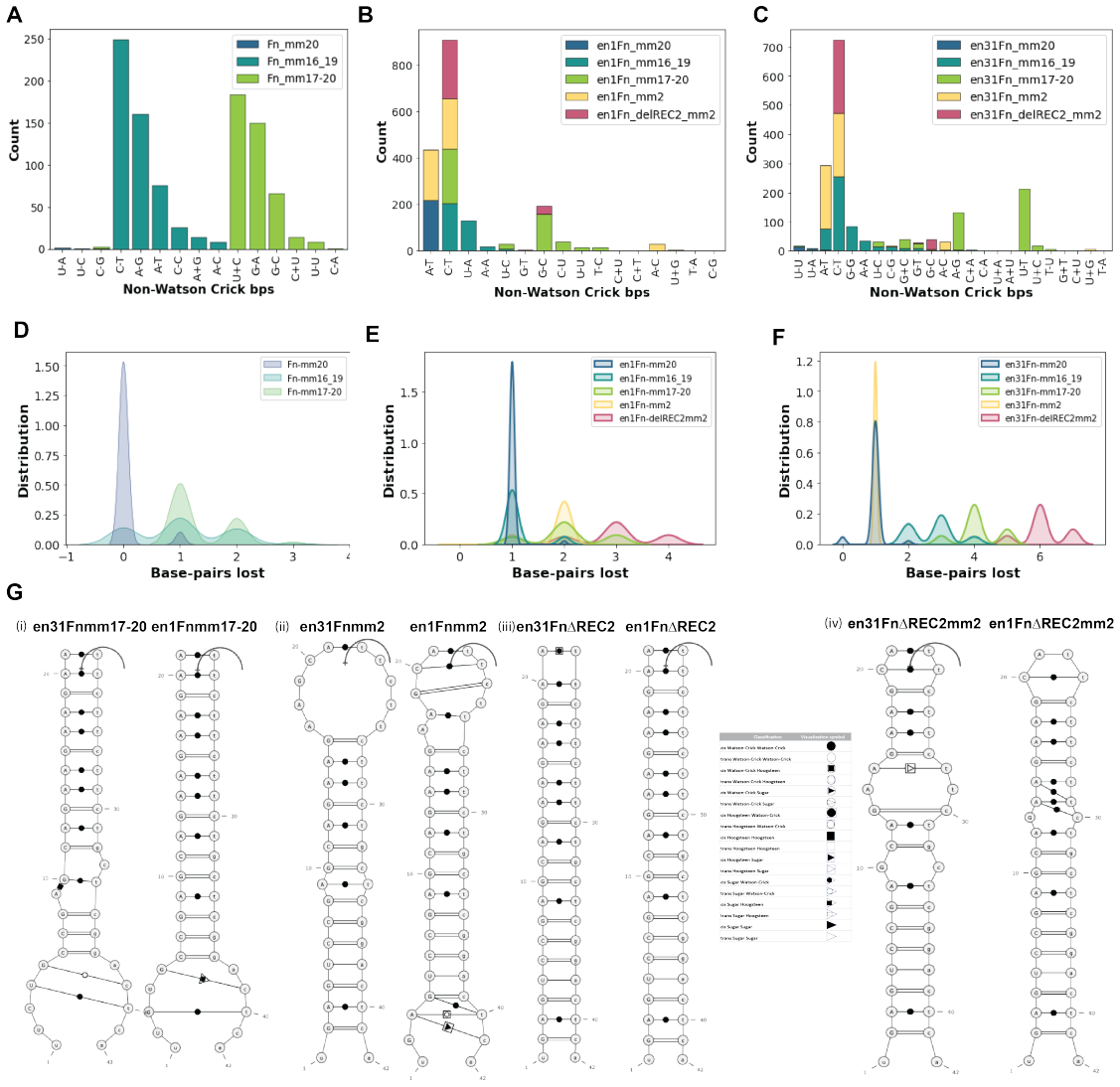


Figure 5.7: Stacked bar plot showing the distribution of non-canonical base pairs found in off-target systems of (i) FnCas9 (ii) en1FnCas9, and (iii) en31FnCas9. B. Distribution plot showing loss of base-pair count in off-target forms of (i) FnCas9 (ii) en1FnCas9, and (iii) en31FnCas9.

is primarily due to its unique non-canonical base-pairing patterns and increased base-pair loss in off-target systems. The broader distribution of base-pair losses and the prevalence of diverse non-canonical pairs, particularly in the four-mismatch system, underscore its higher sensitivity to DNA mismatches. The presence of PAM-distal mismatches disrupts Watson-Crick base pairing but allows non-Watson Crick base-pairing maintaining structural integrity of the duplex, while proximal mismatches lead to more significant structural instability, particularly affecting cleavage efficiency. Clustering analysis indicates that the removal of REC2 mitigates the destabilizing effects of proximal mismatches, enhancing base-pairing integrity. Additionally, analysis of domain dynamics showed that en31FnCas9 exhibits greater fluctuations in the HNH and REC1 domains, especially in critical residues. The MMGBSA calculations further highlighted that while en31FnCas9 exhibits the highest binding affinity for sgRNA-tDNA, its protein-heteroduplex binding is largely mediated by the REC3 domain, with reduced contributions from the HNH and RuvC domains. These findings collectively suggest that the unique structural and dynamic features of en31FnCas9,

particularly the enhanced flexibility and specific interactions within the REC3 domain, contribute to its superior performance.

CHAPTER 6

Expanding FnCas9 Functionality: Comparative Analysis of RNA and DNA Targeting

6.1 Introduction

CRISPR-Cas systems have emerged as powerful tools for RNA targeting, particularly through the use of specific Cas proteins such as Cas13 and type III systems (Csm/Cmr). Cas13, a member of the type VI CRISPR system, is designed to bind and cleave RNA molecules, making it particularly useful for applications in transcriptome engineering, viral interference, and nucleic acid detection. The ability of these systems to selectively target RNA sequences allows for precise manipulation of gene expression without the permanent alterations associated with DNA targeting method Gupta *et al.* (2023); Zhao (2021). The applications of RNA-targeting CRISPR systems are diverse. They include viral diagnostics, where Cas13-based tools can detect viral RNA, providing rapid and sensitive diagnostic capabilities. This has been particularly relevant in the context of emerging infectious diseases Zhao (2021). Additionally, RNA targeting is being explored in cancer therapy, where specific mRNA sequences can be targeted to inhibit oncogenes or modify cancer cell behavior Zhang *et al.* (2022). For instance, the LwCas13a protein has shown promise in targeting mutant KRAS mRNA, achieving significant knockdown efficiency Peterson *et al.* (2022). Furthermore, these systems can be utilized in antibacterial strategies, targeting antimicrobial resistance genes without relying on DNA sequences, thus reducing the risk of resistance development. Moreover, RNA-targeting technologies facilitate therapeutic interventions that do not induce permanent changes to the genome. This is particularly advantageous in treating diseases where traditional gene therapy approaches may pose risks or be ineffective due to difficulties in delivering therapies to specific tissues Brown (2024) The versatility and specificity of CRISPR-Cas RNA-targeting systems hold great potential for advancing both basic research and therapeutic applications across various fields Wan (2019).

Cas12 and Cas Cas13 have been utilized in various gene regulation and therapeutic strategies and have shown promise in applications such as antiviral therapies and RNA knockdown. However, Cas9 also possesses the capability to target RNA Zhang *et al.* (2020*b*). Researchers may prefer Cas9 for RNA cleavage due to several compelling factors. First, Cas9 exhibits dual functionality, enabling manipulation of both DNA and RNA within a single system, thus simplifying experimental design and reducing the need for multiple tools Cong *et al.* (2013*a*). Second, there are extensive established protocols and resources available for Cas9, facilitating easier implementation and optimization of RNA cleavage experiments Jinek *et al.* (2012*a*). Third, Cas9 can offer higher specificity for certain RNA targets compared to Cas12 or Cas13; the ability to design precise

guide RNAs enhances targeting accuracy, which is critical for minimizing off-target effects Mali *et al.* (2013). The well-established structural insights into Cas9 further aid in understanding its mechanism of action and interactions with RNA, thereby improving the design of effective guide RNAs and enhancing targeting efficiency Nishimasu *et al.* (2014*b*). Additionally, the potential for RNA editing applications using Cas9—such as correcting RNA mutations or modifying splicing patterns—adds significant value in therapeutic scenarios where alterations are needed without changing the DNA Zhang *et al.* (2020*a*). Finally, Cas9's compatibility with existing CRISPR systems designed for DNA targeting will allow seamless transition to RNA applications without extensive re-engineering of experimental setups.

The main objective of this work is to compare the domain conformational changes caused by replacing tDNA with tRNA in FnCas9. Specifically, we aim to investigate how tRNA interacts with FnCas9, identify the main interacting residues, determine whether it engages in specific binding or interacts with the same residues as tDNA, and explore whether these residues can be targeted to enhance binding and cleavage efficiency. Additionally, we will assess the binding affinities of tRNA and tDNA with sgRNA and examine the potential role of the WED-PI domain in these interactions. To address these questions, we utilize the FnCas9 ternary complex as a template, removing the ntDNA strand and converting tDNA to tRNA. Atomistic simulations provide a detailed view of molecular interactions at the atomic level, enabling us to capture dynamic conformational changes and binding affinities that traditional methods may overlook. By employing this protocol, we can gain insights into the specific residues involved in tRNA binding and their implications for enhancing Cas9's efficiency in RNA targeting. Conventional molecular dynamics (MD) simulations are performed on both tDNA- and tRNA-bound systems of FnCas9, allowing for a comparative analysis. Studying these interactions is crucial for understanding the mechanistic differences between tRNA and tDNA binding, which can inform strategies for improving Cas9 functionality.

6.2 Materials and Methods

6.2.1 Structural Modelling

Atomistic molecular dynamics simulations have been performed using the X-ray structure of *Francisella novicida* Cas9 (PDB ID: 5B2O) Hirano *et al.* (2016) in complex with RNA and DNA, solved at 1.7 Å resolution. In the case of FnCas9, the nontarget strand (ntDNA) was extended by ten nucleotides, utilizing the ntDNA strand of SpCas9 as a template. The nucleotides on the sgRNA strand of FnCas9 were mutated to target the human EMX1 gene sequence (hEMX1) (5′-GAGUCCGAGCAGAAGAAGAA-3′). This structure was taken a template to model the target-RNA (referred as tRNA) bound structure of FnCas9 where the ntDNA strand was removed and all DNA nucleotides were converted to RNA using mutagenesis module of PyMol **??**.

6.2.2 Classical MD simulations

All constructed model systems were minimized in Chimera Pettersen *et al.* (2004) and used as starting structures for performing simulations. These were immersed in explicit water with the addition of Na+ ions to maintain charge neutrality, creating orthorhombic periodic simulation cells, containing ∼300,000 atoms for the FnCas9 system. Simulations were performed in explicit water with Na+ ions for charge neutrality, using the Amberff14SB Maier *et al.* (2015) force field with specific corrections for DNA and RNA, and the Aqvist model for Mg2+ ions. The systems were solvated in explicit water with Na+ ions for charge balance, creating orthorhombic periodic cells with approximately 300,000 atoms for FnCas9. Equilibration was performed using conventional MD with the Amberff14SBMaier *et al.* (2015) force field, including specific corrections for DNA and RNA, and the Aqvist Aqvist (1990) model for Mg2+ ions. Energy minimization relaxed water molecules and counter-ions, while the system was gradually heated from 0 to 100 K in the NVT ensemble with 50 kcal/mol restraints for 100 ps, followed by an increase to 200 K in the NPT ensemble with 25 kcal/mol restraints for ∼100 ps. Releasing all restraints, the temperature was raised to 300 K in a single 500 ps NPT simulation. The production run was conducted for 50 ns in the NPT ensemble using the GPU version of AMBER22 HMADA *et al.* (2022).

Energy minimization was performed to relax water molecules and counter-ions around the FnCas9 complex before the equilibration steps began. The equilibration consisted of four stages with progressively weaker restraints applied to the protein backbone. In the first equilibration stage, the backbone atoms were restrained with a force constant of 50 kcal/mol, and the system was simulated under the NVT ensemble at 300 K for 10 ns, allowing the water and ions to adapt to the solvated environment. In the second equilibration step, the backbone restraints were reduced to 15 kcal/mol, and the system was again simulated for 10 ns under the same conditions. In the third equilibration step, all restraints were removed, and the system was equilibrated for an additional 10 ns. After equilibration, a gradual heating process was performed. Initially, the system was heated from 0 to 100 K in the NVT ensemble over 100 ps with 50 kcal/mol restraints applied to the system. Then, the temperature was raised to 200 K in the NPT ensemble for 100 ps with 25 kcal/mol restraints. Finally, the system was further heated to 300 K in a single NPT simulation for 500 ps, during which all restraints were removed. Once the system was fully equilibrated, three independent replicate simulations were generated for each FnCas9 complex bound to tDNA and tRNA. Each replicate simulation used the same starting coordinates but different initial velocities to ensure independent sampling of phase space. These production simulations were conducted for 500 ns each, using the NPT ensemble with periodic boundary conditions. The GPU-accelerated version of AMBER22 \cite{hmada2022amber2022} was employed for all runs. A non-bonded interaction cutoff of 8.0 Åwas used, with the Particle Mesh Ewald (PME) method handling long-range electrostatics, and SHAKE constraints were applied to hydrogen bonds, allowing for a 2 fs timestep

6.2.3 MD Simulation Trajectory Analysis

Geometrical properties, including the RMSD of protein backbone atoms (Å) and RNA:DNA hybrid, domain-wise radius of gyration and residue-wise RMSF were also computed for both the Cas9s in tDNA and tRNA bound form. All the analysi were performed for 100-1000ns part of the trajectory. To investigate the dominant motions and conformational distributions of Cas9 domains, the covariance matrix was built and diagonalized using the Pytraj module Roe and Cheatham III (2013). The analysis was performed using the cartesian coordinates for the RuvC, HNH, REC1, REC2, and REC3 domains. The analysis used all frames from all replicates of both Fncas9 bound to tDN and tRNA. To determine key residues which are involved in making tight contacts (within 4 Å) with the susbtrate (tRNA/tDNA) in $>75\%$ frames in the stable part of the trajectory (100-1000 ns). These residue-nucleotide pairs are represented as a network using NetworkX Hagberg *et al.* (2008) package.

6.2.4 MM/GBSA Analysis

The Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method Genheden and Ryde (2015) is a popular technique for determining binding free energies between proteins and ligands. gmx_MMPBSA Valdés-Tresanco *et al.* (2021) was used to calculate the binding free energy of the substrate (tRNA/tDNA) with the FnCas9 RNP complex and for WED and PI domains. These calculations were done for all the replicates of both FnCas9 bound to tRNA and tDNA forn stable part of the trajectory.

6.2.5 Residue-specific RMSF Difference Analysis

To analyze the flexibility of residues interacting with tDNA and tRNA, root mean square fluctuation (RMSF) profiles were computed for residues uniquely interacting with tDNA, those common to both tDNA and tRNA, and those unique to tRNA across all replicates of the FnCas9-tDNA and FnCas9-tRNA complexes. This analysis was performed on the trajectory data from 100 ns to 1000 ns. The RMSF values were calculated using the PyTraj package Roe and Cheatham III (2013). The average RMSF for each residue is presented as a bar plot, with standard deviation included as error bars, providing a visual representation of the variability in residue mobility across simulations. The RMSF difference plot was made subtracting the average RMSF values for the tRNA-bound system from those of the tDNA-bound system. Positive values indicate higher flexibility in the tDNA-bound system (represented by blue bars), while negative values suggest greater flexibility in the tRNA-bound system (represented by red bars).

6.2.6 Distance Analysis

Distance analyses were conducted for residues uniquely interacting with tDNA and tRNA across all replicates of the FnCas9 systems. The analysis focused on calculating the distances between these residues and their nearest nucleotide. Using the cpptraj package Roe and Cheatham (2013), distances were computed for the entire trajectory of each replicate for both the FnCas9-tRNA and FnCas9-tDNA complexes.

6.3 Results and Discussions

6.3.1 Structural Stability and Flexibility of FnCas9 Complexes with tDNA and tRNA

To compare the structural stability of modelled tRNA bound FnCas9 complex with tDNA bound FnCas9, the protein-backbone RMSD of the forms of Cas9 were evalused and compared for all replicate systems (shown in Figure 6.1A and C). The RMSD plots depict the overall stability of both tDNA- and tRNA-bound complexes over the simulation time, highlighting noticeable differences in their dynamic behavior. The tRNA-bound complex exhibits a relatively higher RMSD, which stablised by the end of the simulation for the replicates suggesting greater flexibility or structural deviation in comparison to the tDNA-bound complex. Overall, the RMSD exhibited by replicates of both the systems have shown a similar overlap suggesting a stabilisation in the system and the higher deviations in tRNA bound complex indicates more dynamic changes in FnCas9, possibly due to different interaction patterns.

To determine the domain or regions of FnCas9 exhibiting structural changes due to tRNA binding, the residue-wise RMSF calculations were done for all the replicates of tRNA and tDNA bound FnCas9 (shown in Figure 6.1B and 6.1D). The RMSF plots provide insights into the per-residue flexibility of various FnCas9 domains. The domain-wise RMSF (shown in Figure 6.1B (ii) and 6.1D(ii)) revealed REC2, RuvC and PI domains are more dynamic than other domains in both the forms of Cas9. The REC2 domains has shown higher fluctuations in tRNA bound state in all the replicates as compared to tDNA bound form indicating that tRNA binding results in distinct flexibility patterns inducing greater conformational shifts in FnCas9, especially in the REC2 and PI domains.

6.3.2 Domain-Specific Motions and conformational changes in tRNA vs. tDNA Systems

Principal component analysis (PCA) was employed to capture the major conformational variations in individual domains of FnCas9 between tRNA- and tDNA-bound systems (shown in Figure 6.2).

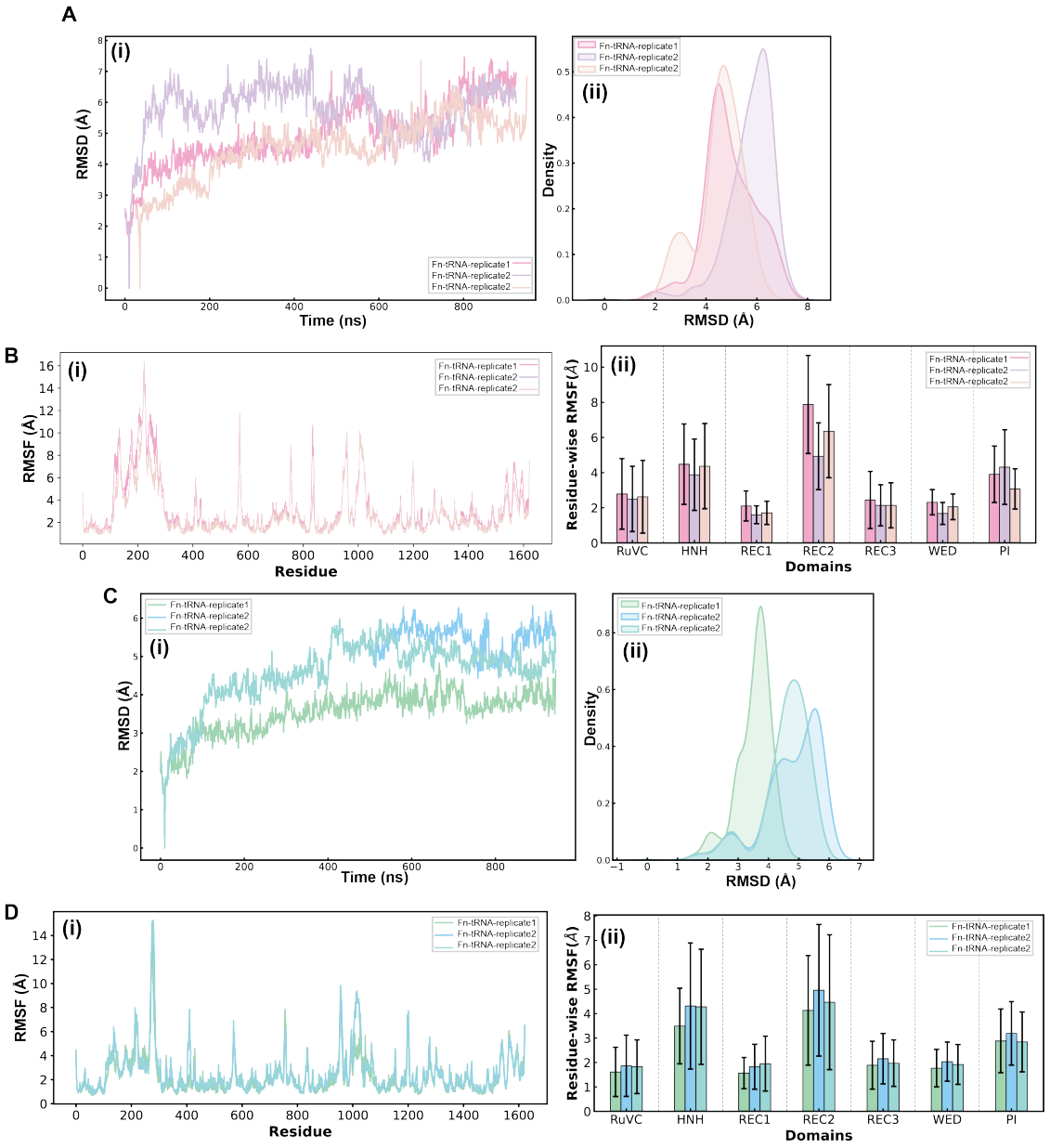


Figure 6.1: (A) RMSD plot (Angstorms) of protein backbone atoms over time for FnCas9-sgRNA-tRNA bound form (i) Line plot showing RMSD for three replicates. (ii) Probability distributions of RMSDs in all three replicates. (B) RMSF comparison in all three replicates of FnCas9-sgRNA-tRNA bound forms (i) Residue-wise RMSF line plot (Angstorms) (ii) Domain-wise-RMSF bar-plots in all three replicates. (C) RMSD plot (Angstorms) of protein backbone atoms over time for FnCas9-sgRNA-tDNA bound form (i) Line plot showing RMSD for three replicates. (ii) Probability distributions of RMSDs in all three replicates. (D) RMSF comparison in all three replicates of FnCas9-sgRNA-tDNA bound forms (i) Residue-wise RMSF line plot (Angstorms) (ii) Domain-wise-RMSF bar-plots in all three replicates.

This analysis allowed us to compare the structural deviations of different replicates, revealing the primary modes of motion and flexibility. The RuvC domain exhibits more pronounced conformational shifts in the tRNA-bound system than in the tDNA-bound system (shown in Figure 6.2A). This is evident from the larger spread observed in the PCA plot for tRNA replicates, suggesting that the RuvC domain undergoes greater variability in its conformational space. These results in-

dicate that tRNA binding introduces greater flexibility in the RuvC domain, which might affect its role in DNA/RNA cleavage and the overall stability of the complex. Furthermore, the RMSF values for the RuvC domain in the tRNA-bound system (shown in Figure 6.1B) are consistently higher across multiple replicates compared to the tDNA-bound system, confirming that tRNA induces greater fluctuations and structural deviations in this domain. This increased flexibility may play a role in how FnCas9 processes different nucleic acid substrates. While all the other domains have shown a similar conformational deviation in both the systems, where REC2 (Figure 6.2E) domain has shown highest conformational change followed by HNH domain in both tRNA and tDNA bound system. This suggests that REC2 is intrinsically flexible and undergoes substantial conformational rearrangements irrespective of the nucleotide substrate. The WED and PI domains exhibit relatively small shifts in their conformational distributions compared to the highly dynamic RuvC and REC2 domains, particularly in the context of tRNA binding.

On comparing the structural compactness of these domains, similar observations were made where RuvC was found to higher radius of gyration suggesting a comparatively less compact structure than the one found in tDNA bound system, suggesting that tRNA binding affects the arrangement of RuvC domain because of distinct interaction pattern in FnCas9.

6.3.3 Binding Affinity Differences and the Role of WED and PI Domains

MM/GBSA calculations were conducted for both tRNA- and tDNA-bound FnCas9 systems to evaluate the binding affinities of these substrates with the entire ribonucleoprotein (RNP) complex. The binding affinity for tDNA (20 bp) ranged from -224.05 to -249.54 kcal/mol, while the binding affinity for tRNA (20 bp) varied between -191.93 and -211.77 kcal/mol. These results indicate that tDNA exhibits a stronger binding affinity to the RNP complex compared to tRNA across all replicates (Table 6.1). Although the difference in binding energy is modest, it consistently supports the conclusion that tDNA binds more tightly to FnCas9 than tRNA. To further investigate the role of the WED and PI domains in substrate binding and recognition, we also calculated the binding affinities of these domains with the full tRNA and tDNA strands. Interestingly, the WED and PI domains demonstrated a higher binding affinity for the tRNA strand than for the tDNA strand, with values ranging from -71.3 to -83.5 kcal/mol for the tDNA replicates and from -110.1 to -131.4 kcal/mol for the tRNA replicates. This suggests that while tRNA binding may induce conformational changes in FnCas9 that reduce its overall affinity for the RNP complex, it simultaneously enhances the affinity of the WED and PI domains for tRNA. These findings underscore the nuanced interactions between FnCas9 and its substrates, highlighting that different RNA forms can influence binding dynamics and conformational states within the protein complex. Such insights are critical for understanding substrate recognition mechanisms and optimizing Cas9 functionality in RNA-targeting applications.

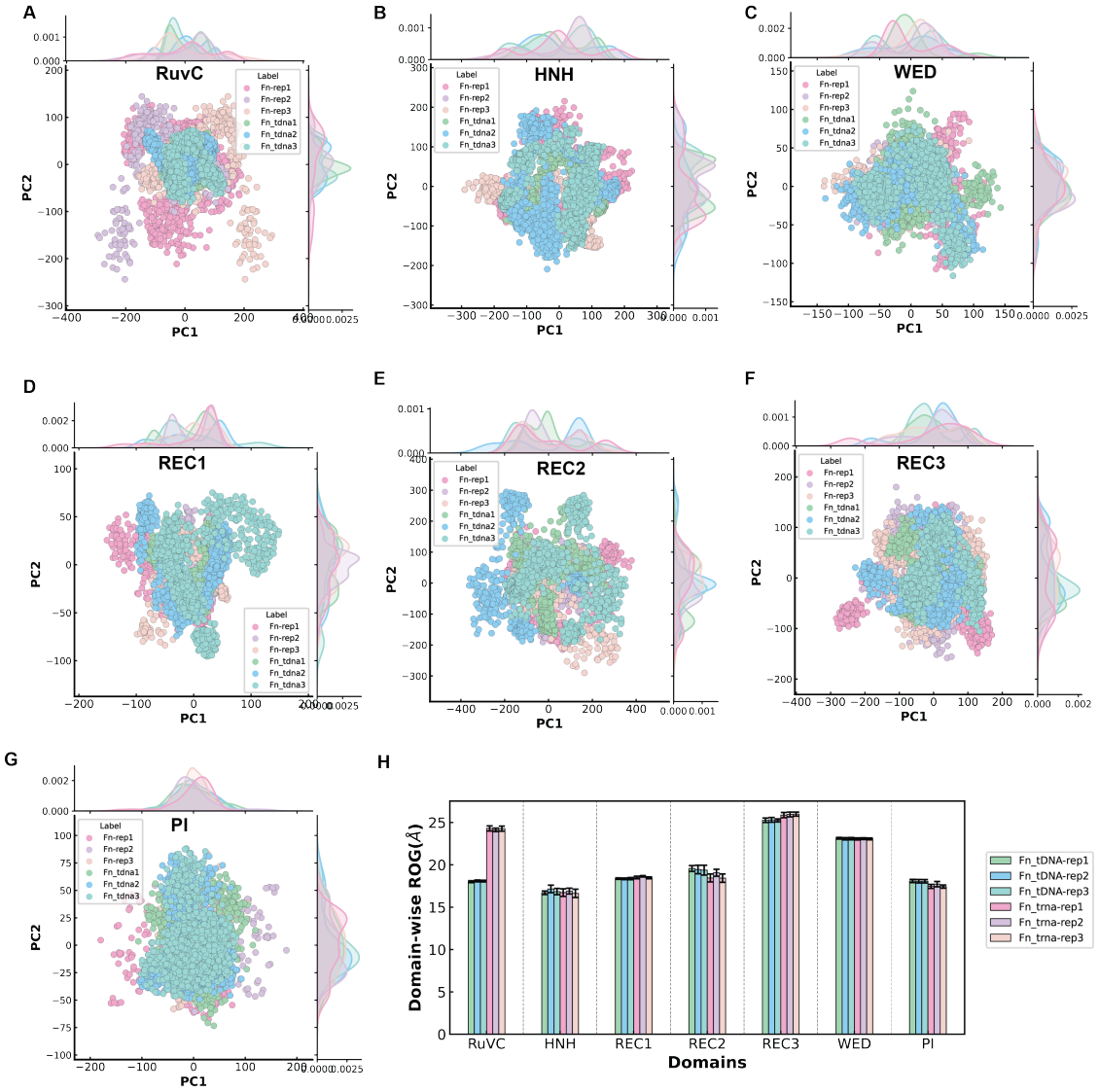


Figure 6.2: Principal Component Analysis (PCA) graphs illustrating the domain-wise conformational distributions of the (A) RuvC, (B) HNH, (C) WED, (D) REC1, (E) REC2, (F) REC3, and (G) PI domains, (H) Domain-wise radius of gyration in all the replicates of FnCas9 bound to tRNA (shades of pink) and tDNA (shades of blue),

6.3.4 Residue-Specific Binding Interactions: A Comparative Contact Map Analysis

To understand how tRNA strand interacts with the FnCas9, is it different from how tDNA interacts with it, what are the critical residues common to both the substrates and specific to each substrates, the contact-analysis was performed. In this analysis, all those residues which maintained a tight contact (< 4.5 Å) in more than 75% of the frames from 100-1000ns part of the trajectory were considered. All the interacting residues (both substrate and protein) for each replicate system is shown as network where residues circles are colored depending on the domain to which they belong (Figure 6.3).

In the tDNA-bound systems (Figure 6.3A) , contacts made by residues from REC1 and REC3

System	Protein-substrate (20 bp) Energy (Kcal/mol)	WED_PI-substrate (entire stretch) Energy (Kcal/mol)
Fn-native–tRNA (rep1)	-211.77 ± 7.34	-131.45 ± 7.25
Fn-native–tRNA (rep2)	-210.48 ± 3.34	-124.85 ± 6.43
Fn-native–tRNA (rep3)	-191.93 ± 5.43	-110.09 ± 5.46
Fn-native–tDNA (rep1)	-249.54 ± 4.48	-83.45 ± 4.45
Fn-native–tDNA (rep2)	-224.15 ± 5.36	-76.88 ± 9.10
Fn-native–tDNA (rep3)	-224.05 ± 3.16	-71.28 ± 8.86

Table 6.1: Binding energy comparison between FnCas9 and tRNA/tDNA substrates: The table shows the binding energies (kcal/mol) of FnCas9 with tRNA and tDNA substrates for a 20 bp region, as well as the interaction energies between the WED-PI domain and the substrate across three replicates. The data reveal a stronger binding affinity for the tDNA-bound complex compared to the tRNA-bound complex, with mean values reported ± standard deviation.

are more prominent across all replicates with 1-2 residues from HNH domain and L1-L2 loops. REC1 and REC3 have strong conservation of contacts across replicates, with 16 residues common in all three which were : 106, 450, 451, 452, 453, 458 (from REC1 domain), 663, 789, 794, 798, 809, 810, 845, 848, 856 (from REC3 domain), and 1042 (from HNH domain), suggesting their role in tDNA recognition. In tRNA-bound systems (Figure 6.3B), residues from REC1, REC3, WED domains, and L2-loop were found to be interacting in all the replicates. However, the noticeable increase in contacts within the WED domain and L2 loop in these systems indicates the conformational changes in these regions are associated with the binding of tRNA substrates. The RuvC domain shows fewer contacts compared to the tDNA-bound systems, suggesting that this domain may have less direct involvement in tRNA binding.

In terms of overlap, the Venn diagram shows 16 common contacts across all tDNA replicates (Figure 6.3C(i)) , suggesting a high degree of structural conservation in these regions. The tRNA replicate systems share 14 common interacting residues (Figure 6.3C(ii)) which were: 450, 451, 452, 453, 808, 809, 810, 789, 663, 857, 1082, 1083, 1084, and 858. While among these sixteen and fourteen common residues, eight(Figure 6.3C(iii)) were common to both tRNA and tDNA bound systems which were: 450, 451, 452, 453, 809, 810, 789, and 663. While both systems share eight common residues, tDNA systems display 8 unique contact residues and tRNA systems show 6 unique contacts. This indicates distinct structural behaviors between the two systems, with tDNA systems exhibiting more conserved interactions across its replicates. This divergence likely reflects the differences in substrate recognition and binding dynamics for tDNA versus tRNA. These findings highlight the differential structural adaptation of FnCas9 when bound to tRNA versus tDNA, which may influence its cleavage activity and specificity for these substrates.

6.3.5 Analysis of Residue-Specific Flexibility and Substrate-Specific Binding

The contact analysis identified eight unique residues interacting with the tDNA strand of FnCas9 (106, 458, 794, 798 , 845, 848, 856, and 1042) while the tRNA strand formed six unique contacts

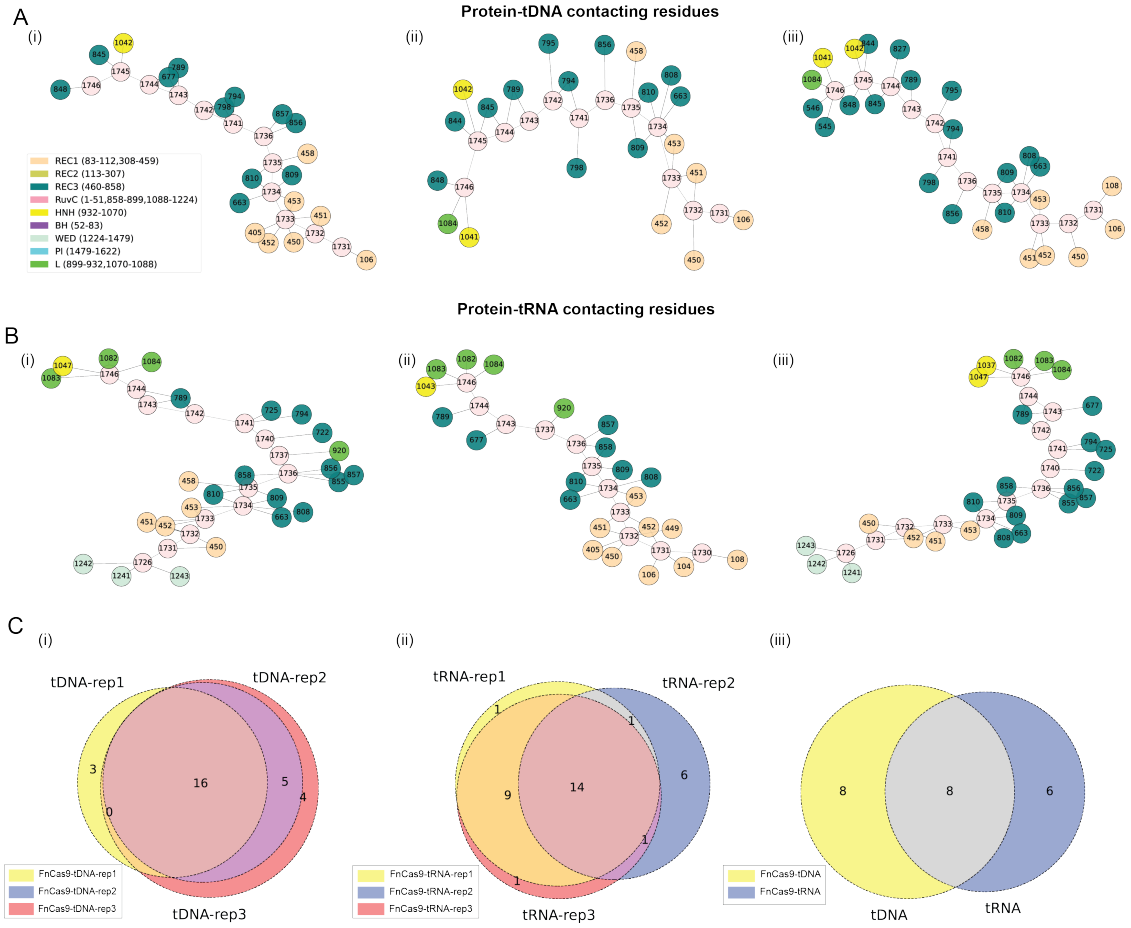


Figure 6.3: Domain-wise contact networks, highlighting key residues within different domains such as REC1, REC2, RuvC, HNH, WED, PI, and L1 and L2 loops. Residue-residue contact maps for different replicates of (A) FnCas9 bound to tDNA, (B) FnCas9 bound to tRNA, and (C) Venn-diagram showing number of common/unique interacting residues in all three replicates of (i) FnCas9 bound to tDNA (ii) FnCas9 bound to tRNA, and (iii) Both FnCas9 bound to tDNA and tRNA.

(808, 857, 858, 1082, 1083, and 1084). Additionally, eight residues were found to be common interaction points between both the tDNA and tRNA substrates (Figure 6.4A). A comparison of the flexibility of these substrate-specific residues was carried out by calculating their residue-wise root mean square fluctuation (RMSF) values, as variations in flexibility could explain differences in contact retention between the substrates. The RMSF profiles were computed for residues uniquely interacting with tDNA (Figure 6.4B), common to both tDNA and tRNA (Figure 6.4C), and unique to tRNA (Figure 6.4D) across all replicates for both FnCas9-tDNA and FnCas9-tRNA complexes. To further explore the differences in mobility, the average RMSF values for the tRNA-bound system were subtracted from those of the tDNA-bound system for each residue. Positive values (blue bars) indicate higher flexibility in the tDNA-bound system, whereas negative values (red bars) suggest higher flexibility in the tRNA-bound system.

In the case of residues unique to the tDNA-bound system (Figure 6.4E), the RMSF differences revealed greater flexibility in the tRNA-bound system, with Phe1042 showing the most pronounced increase in mobility. This suggests that this residue forms more stable interactions in

the tDNA-bound system compared to the tRNA-bound system. Residues common to both tDNA and tRNA (Figure 6.4F) showed mixed flexibility patterns, with residues 450-453 and 663 being more mobile in the tDNA-bound system, while residues 789, 809, and 810 exhibited higher mobility in the tRNA-bound system. Interestingly, residues unique to the tRNA-bound system showed increased flexibility in the tDNA-bound system (Figure 6.4G). Specifically, residue 858 demonstrated enhanced stability in the tRNA-bound system, reinforcing its role in stable tRNA interactions. Collectively, these results suggest that residues specifically involved in tRNA binding tend to be more flexible in the tDNA-bound system, and vice-versa. This indicates that interactions with tDNA are more dynamic compared to tRNA, highlighting the distinct conformational and interaction landscapes of FnCas9 with these two substrates.

6.3.6 Dynamic Distance Fluctuations of Key Binding Residues

The residue-wise RMSF analysis highlights that residues 106, 458, 794, 798, 845, 848, 856, and 1042 form stable interactions with tDNA, while residues 808, 857, 858, 1082, 1083, and 1084 demonstrate stable contacts with tRNA. This suggests a distinct interaction landscape between the FnCas9 complex and tRNA compared to tDNA.

To further substantiate these observations, distance analyses were performed for the aforementioned residues across all replicates of the Cas9 systems bound to tRNA and tDNA. This allowed for the examination of changes in the distances of these residues over time, particularly to assess if any of them moved significantly apart, leading to a potential loss of interaction. The analysis revealed that residues 857, 858, and 1082 maintained stable contacts throughout the simulations, especially in the tRNA-bound system. However, in the tDNA-bound system, these residues exhibited more dynamic behavior, frequently moving farther from the tDNA strand, as observed across all replicates (Figure 6.5).

Similarly, residues 845, 848, and 1042, which form stable interactions in the tDNA-bound system, were observed to lose contact with the tRNA strand, moving further apart in the tRNA-bound simulations (Figure 6.6). These findings further confirm the differential binding preferences of these residues, with 857, 858, and 1082 showing a preference for tRNA binding, and residues 845, 848, and 1042 displaying preferential binding to tDNA.

6.4 Conclusions

This study investigates the structural and binding dynamics of FnCas9 in complex with tRNA and tDNA substrates through molecular dynamics simulations, contact analysis, MM/GBSA binding free energy calculations, and principal component analysis (PCA). Our findings reveal distinct interaction networks for each substrate. In tDNA-bound systems, residues from the REC1 and REC3 domains—specifically 106, 450, 451, 452, 453, 458, 663, 789, 794, 798, 809, 810, 845, 848, 856,

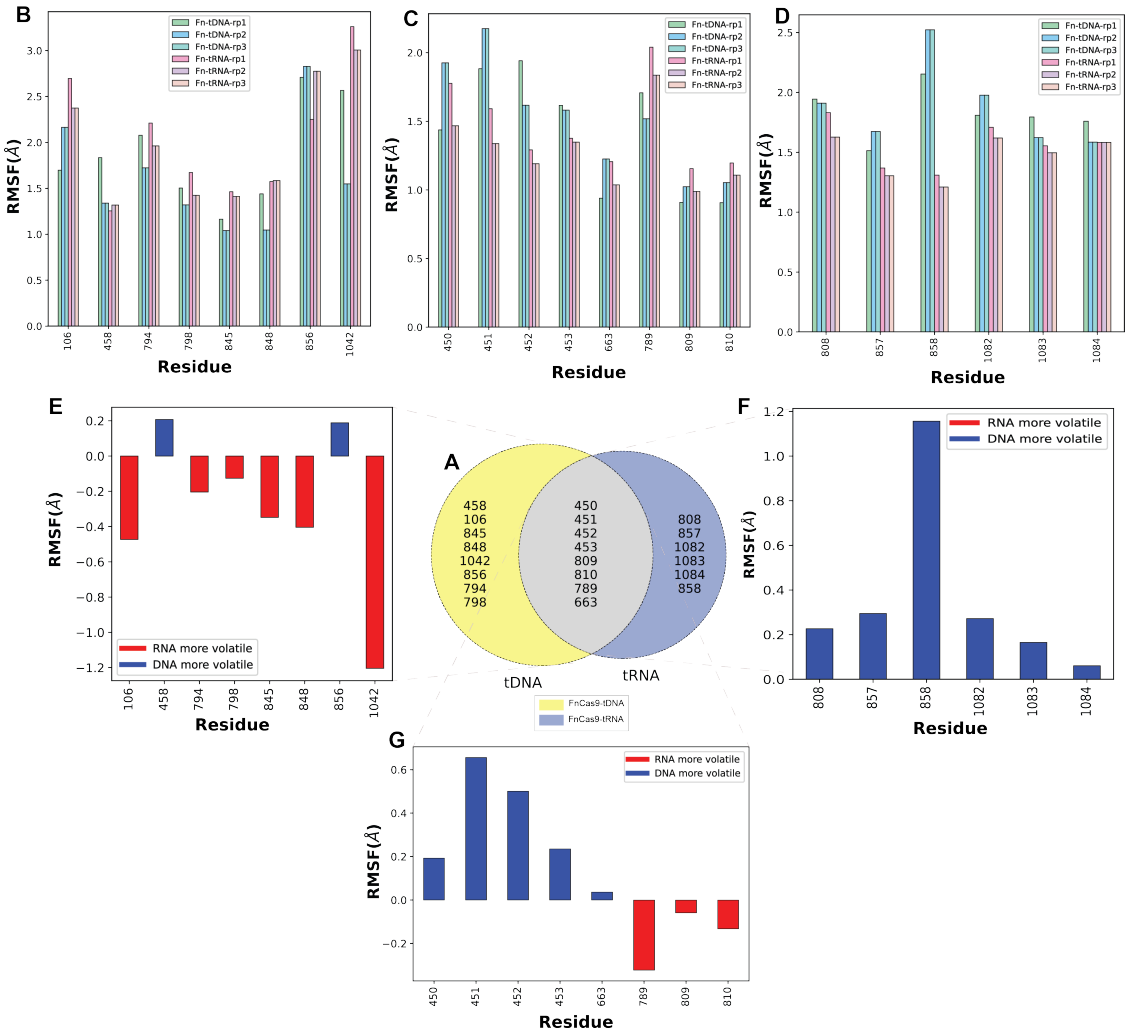


Figure 6.4: RMSF Analysis for Common and Unique Binding Residues in FnCas9-tDNA and Fncas9-tRNA systems: A. The central Venn diagram identifies residues common to both complexes (gray overlap), residues unique to tDNA binding (yellow), and residues unique to tRNA binding (blue). RMSF across three replicates (rep1, rep2, rep3) for residues involved in B. Specific binding to tDNA (left), C. Common in both tDNA and tRNA systems, and D. Specific binding to tRNA (right). RMSF difference plots made by subtracting RMSF of residues in tRNA from tDNA system. Positive value (blue bar) means higher RMSF in tDNA system, and vice-versa (red bar). E. RMSF difference plot for residues unique to tDNA. F. RMSF difference plot for residues unique to tRNA, G. RMSF difference plot for residues common in both tDNA and tRNA.

and 1042 (HNH)—are prominently involved across all replicates, indicating their central role in tDNA recognition. Conversely, in tRNA-bound systems, contacts shift to include residues from the REC1, REC3, WED domains, and the L2 loop, with fewer contacts observed in the RuvC domain. This suggests a reduced involvement of the RuvC domain in tRNA binding. Unique residues associated with tRNA include 857, 858, and 1082–1084, reflecting specific structural interactions required for RNA binding. Notably, eight residues—450, 451, 452, 453, 789, 809, 810, and 663—are common to both substrates, highlighting shared recognition elements while underscoring distinct interaction patterns. The MM/GBSA analysis indicates that while tDNA ex-

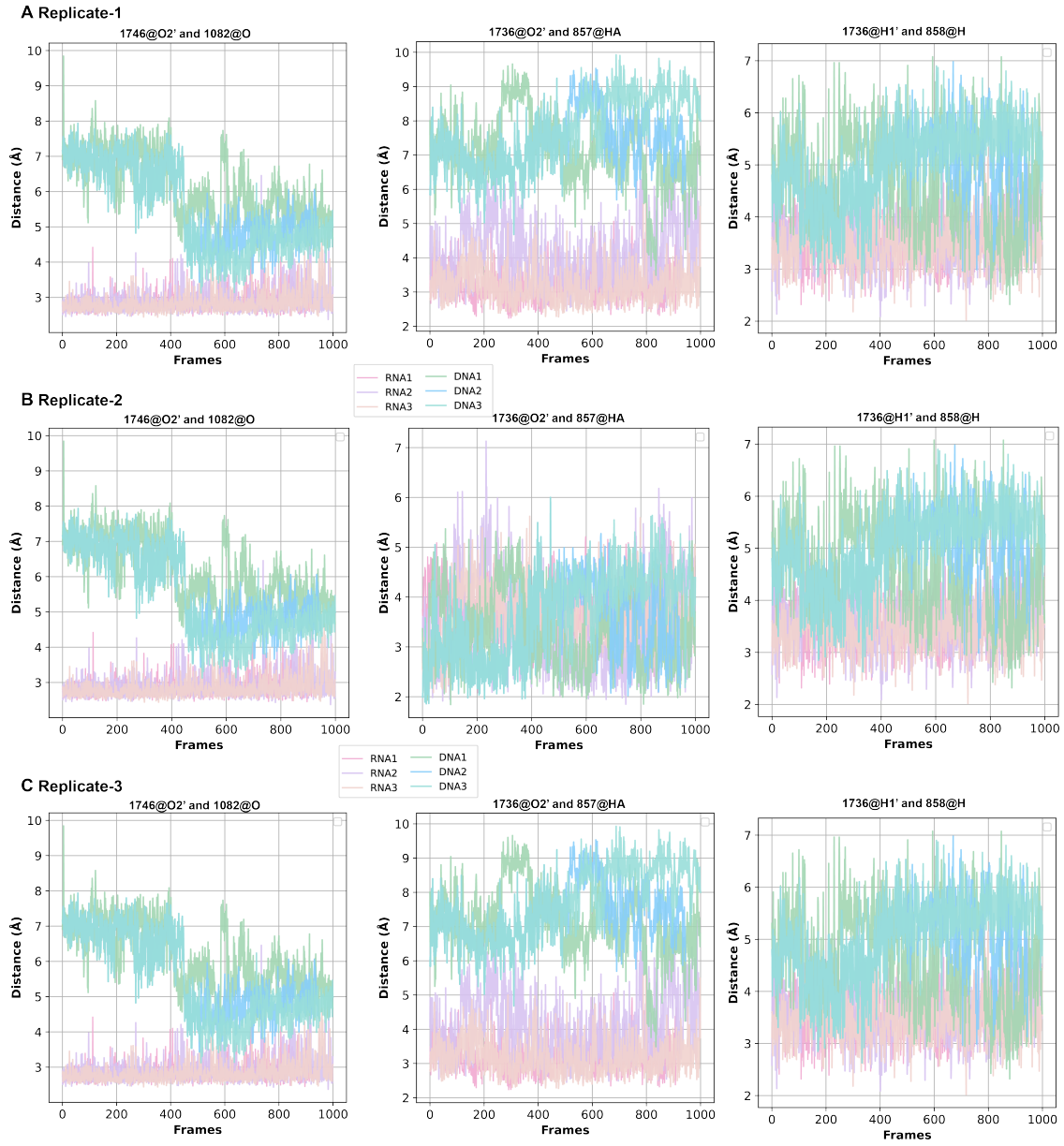


Figure 6.5: Distance plot for residues (1082, 857, and 858) unique to tRNA bound system which went farther away in tDNA system for (A) Replicate 1, (B) Replicate 2, and (C) Replicate 3 of both tRNA and tDNA bound systems.

hibits a stronger overall binding affinity to FnCas9 compared to tRNA, the WED and PI domains show a preference for tRNA binding. This suggests that tRNA binding may induce conformational changes that modulate FnCas9's affinity across different regions of the protein. PCA further emphasizes differences in flexibility and conformational dynamics between the two substrate-bound systems; specifically, the RuvC domain exhibits increased fluctuations in the tRNA-bound system. This increased flexibility confirms greater structural variability when bound to tRNA and may influence cleavage activity. The RMSF analysis aligns with these observations; residues 857, 858, 1082, and 1083 show higher RMSF values in tRNA-bound systems. In contrast, the tDNA-bound systems demonstrate lower RMSF values in the RuvC domain, indicating a more stable structure during DNA binding.

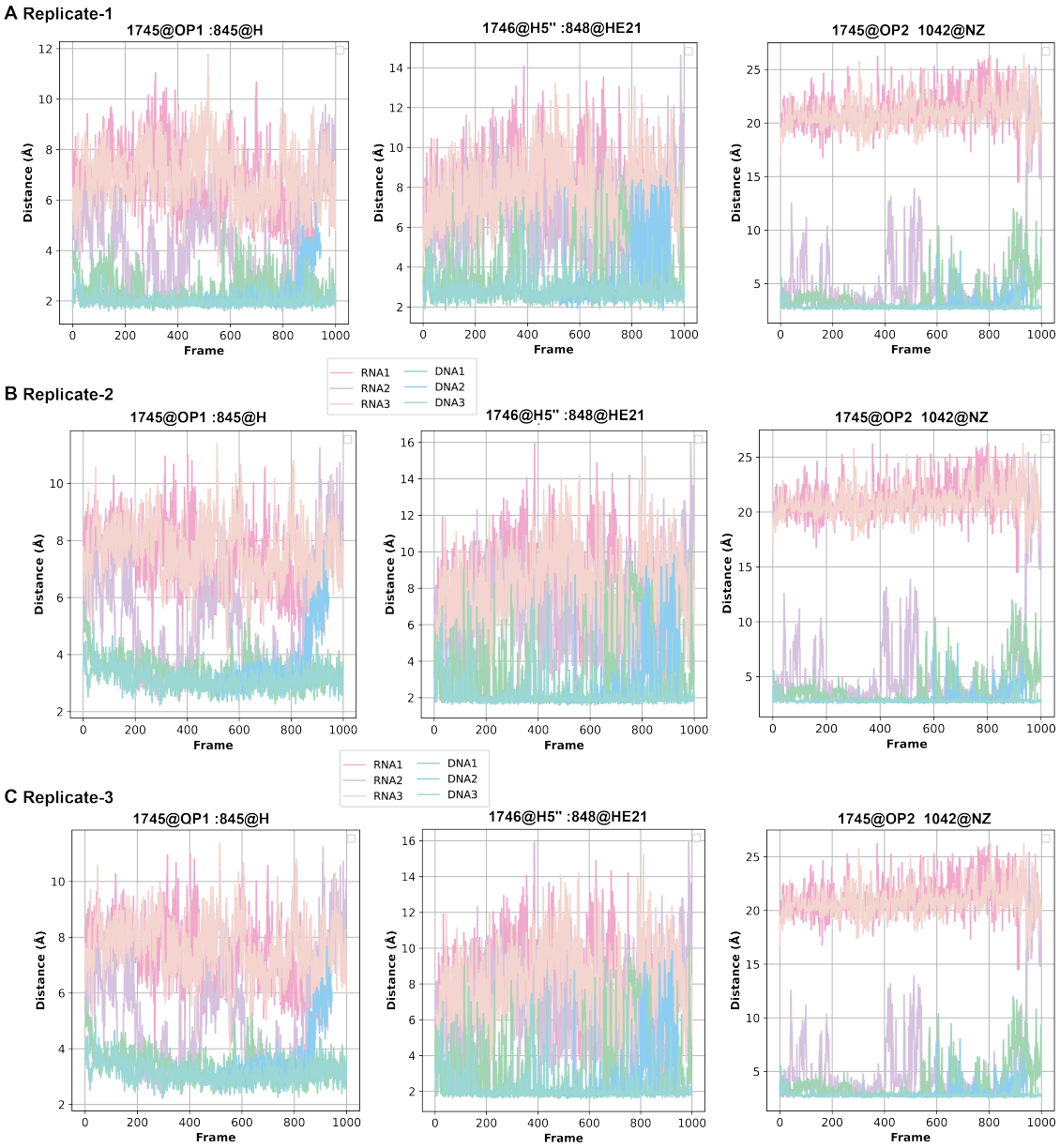


Figure 6.6: Distance plot for residues (845, 848, and 1042) unique to tDNA bound system which went farther away in tRNA system for (A) Replicate 1, (B) Replicate 2, and (C) Replicate 3 of both tRNA and tDNA bound systems.)

Distance-based analysis supports these findings by showing greater distance variations between key residues in the WED and RuvC domains for tRNA-bound systems compared to those bound to tDNA. This implies that tRNA binding induces more dynamic conformational shifts than those observed with tDNA. Overall, our study elucidates key differences in the structural and dynamic behavior of FnCas9 when interacting with tRNA versus tDNA. These insights could be leveraged for optimizing CRISPR-Cas systems for RNA-targeting applications and inform the design of Cas9 variants with enhanced specificity or activity for targeted editing.

CHAPTER 7

Conclusions and Future Work

This study offers a comprehensive examination of the structural and dynamic properties of two Cas9 orthologs, SpCas9 and FnCas9, highlighting their distinct functional characteristics, particularly in terms of gRNA binding, specificity, and cleavage mechanisms. The findings provide a deeper understanding of the molecular dynamics governing Cas9 function and open up new avenues for optimizing Cas9-based genome editing.

7.0.1 Summary of Major Conclusions

Objective 1: Comparative structural and dynamics study of free and gRNA-bound FnCas9 and SpCas9 proteins The aim of this study was to explore the structural and dynamic differences between FnCas9 and SpCas9, focusing on how gRNA binding impacts their conformational landscapes and domain organization. Despite being homologous, these Cas9 orthologs exhibit distinct behaviors and cleavage mechanisms when targeting DNA sequences. To understand these distinctions, Molecular Dynamics (MD) simulations were performed on both the apo (unbound) and gRNA-bound states of each protein. The analysis placed particular emphasis on electrostatic interactions and critical structural elements, such as the bridge-helix, to uncover their roles in stabilizing FnCas9 in a cleavage-competent state. Principal Component Analysis (PCA) and Free Energy Landscape (FEL) mapping were employed to investigate protein dynamics, flexibility, and structural rearrangements that might explain FnCas9's higher specificity and reduced off-target effects.

The results revealed that gRNA binding stabilizes both SpCas9 and FnCas9, but it induces distinct structural and dynamic changes in each protein. The binding of gRNA decreased structural compactness in both proteins, suggesting an opening of the structure and exposure of hydrophobic residues. FnCas9 exhibited greater flexibility in its apo form compared to its gRNA-bound state, while SpCas9 showed similar flexibility in both forms. Residue-specific fluctuations highlighted differences in dynamics: in SpCas9, the REC2-REC3 domains (residues 218–233, 727–791) and HNH domain (950–998, 1004–1065) showed high fluctuations in the gRNA-free state, while in FnCas9, the PI domain (1565–1572) displayed significant mobility in the gRNA-bound form. Domain rearrangements triggered by gRNA binding were observed, including movements between HNH-REC3-REC1 and PI-WED-RuvC domains in FnCas9, and between HNH-REC2, CTD-Topo, and REC3-RuvC domains in SpCas9. These dynamic shifts correlated with changes in flexibility and the radius of gyration, particularly in FnCas9, which exhibited larger domain deviations compared to SpCas9. The free-energy landscape analysis revealed that gRNA-bound forms in both proteins achieved more stable structural states compared to their apo forms. Notably, the

MM/GBSA binding energy analysis showed that gRNA is more tightly bound in FnCas9 than in SpCas9, with electrostatic energy (EEL) being the dominant contributor in FnCas9, while polar solvation energy (EGB) played a key role in SpCas9. Residue-specific binding energy decomposition identified key contributors to gRNA binding: Phe105 in SpCas9 and Arg455, Arg401, Arg55, Arg62, and Arg63 in FnCas9. Three of these residues (R55, R62, R63) belong to the bridge-helix of FnCas9, highlighting its critical role in gRNA recognition and binding. Conservation analysis further revealed that the HNH domain of SpCas9, which is highly conserved, exhibited low dynamic residue percentages, while the least conserved REC2 domain showed higher dynamics. In FnCas9, the PI domain's increased coil percentage upon gRNA binding suggested structural transitions contributing to its functional dynamics. These findings suggest that the structural and energetic factors that differentiate FnCas9 and SpCas9. FnCas9's tighter gRNA binding and dynamic rearrangements likely contribute to its higher specificity and reduced off-target effects.

Objective 2: Deciphering Cas9 Specificity: Role of Domain Dynamics and RNA:DNA Hybrid Interactions Revealed through Machine Learning and Accelerated Molecular Simulations

This study investigated the mechanisms underlying FnCas9's enhanced specificity compared to SpCas9, focusing on the effects of base-pair mismatches in the RNA:DNA hybrid. To mimic off-target conditions, PAM-distal mismatches were introduced, and their impact on heteroduplex dynamics and protein responses was analyzed using advanced computational methods. Techniques such as accelerated molecular dynamics (GaMD), Principal Component Analysis (PCA), Markov State Modeling (MSM), unsupervised clustering (UMAP), and binding affinity calculations via MM/GBSA were employed to examine the conformational dynamics, domain interactions, and signal transmission mechanisms in both Cas9 orthologs. These comprehensive analyses shed light on the structural factors that contribute to the observed differences in specificity between FnCas9 and SpCas9.

The results revealed that FnCas9 exhibited fewer conformational changes in the heteroduplex in the presence of base-pair mismatches compared to SpCas9, particularly with pyrimidine-pyrimidine mismatches (C-T/U-C/T). PCA and MSM analyses showed that FnCas9 maintained restricted conformational shifts even under mismatch conditions, indicating that its cleavage mechanism relies more on domain rearrangements than on hybrid shape. UMAP clustering further demonstrated that FnCas9 could better distinguish between on-target and off-target forms by combining hybrid and protein features, suggesting that FnCas9's specificity is driven by interactions between the protein domains and the hybrid, rather than the hybrid's intrinsic shape. In contrast, SpCas9 exhibited greater reliance on hybrid conformational states for mismatch discrimination. Domain interactions played a pivotal role in specificity. In FnCas9, REC1-REC3 contacts with the RNA:DNA hybrid were critical for mismatch discrimination, while SpCas9 relied more heavily on HNH-RuvC-hybrid contacts. MM/GBSA calculations revealed distinct binding affinities: in FnCas9, the REC lobe showed stronger binding to the RNA:DNA hybrid, whereas SpCas9 exhibited higher binding affinity in the NUC lobe. These findings suggest that mismatches differentially affect REC-hybrid binding in FnCas9 and NUC-hybrid binding in SpCas9. Furthermore, critical

residues in the REC and NUC lobes were identified as key contributors to heteroduplex binding, with potential implications for site-directed mutagenesis to enhance specificity. The analysis also highlighted differences in information transmission pathways between catalytic and recognition site. In SpCas9, allosteric information transfer involved the REC3, REC2, and HNH domains, whereas in FnCas9, signal transfer occurred directly from REC3 to HNH, bypassing REC2. This shorter information pathway in FnCas9 likely enhances its specificity by facilitating efficient recognition of mismatches and preventing cleavage of off-target sequences. Interestingly, FnCas9 exhibited greater conformational deviations in its protein domains than SpCas9, suggesting that its cleavage mechanism depends more on domain rearrangements than the structural rigidity of the RNA:DNA hybrid. This was further supported by machine learning-based classification, which showed that FnCas9's heteroduplex states were influenced by intricate domain-hybrid interactions, while SpCas9 relied more on intrinsic hybrid features.. The findings from this study provide critical insights into the mechanisms of mismatch recognition and specificity in Cas9 systems, paving the way for future experimental validations and targeted engineering of high-fidelity variants for improved genome editing applications.

Objective 3: Structural Insights into Mismatch Recognition by Engineered FnCas9 Variants

Engineered FnCas9 variants (en1, en15, and en31) were developed by modifying specific structural features, such as the WED-PI domain and phosphate-lock loop (PLL), without compromising intrinsic specificity. The engineered en31 variant, in particular, demonstrates robust cellular editing, higher cleavage efficiency, and compatibility with extended (x-) and super-extended (sx-) gRNAs, enabling broader targeting capabilities and improved single-base specificity. These attributes make en31 a promising candidate for advancing genome editing technologies. The objective of the study was to investigate the structural and dynamic features of engineered FnCas9 variants and to explore the impact of specific mutations and domain modifications on their efficiency and specificity. Molecular dynamics (MD) simulations were conducted to analyze the behavior of en1, en15, and en31 in both their native and base-pair mismatch states. Comparative analyses with truncated variants lacking the REC2 domain further aimed to uncover the role of REC2 in stabilizing protein-DNA complexes and how its removal influences heteroduplex dynamics and specificity. A particular focus was placed on the effects of PAM-proximal and PAM-distal mismatches on heteroduplex stability, binding affinity, and cleavage efficiency, providing insights into the underlying mechanisms of mismatch recognition.

The results of the study highlighted unique structural and dynamic features behind the enhanced performance of the engineered variants. En31FnCas9 exhibited increased sensitivity to mismatches, particularly in off-target systems, due to the presence of non-canonical base-pairing patterns and a broader distribution of base-pair losses. Notably, PAM-distal mismatches allowed non-Watson-Crick base pairing, preserving the structural integrity of the duplex, while PAM-proximal mismatches caused significant destabilization, leading to reduced cleavage efficiency. Clustering analysis revealed that the removal of the REC2 domain mitigated the destabilizing effects of proximal mismatches, enhancing base-pairing integrity and potentially contributing to

improved specificity. Additionally, domain dynamics analyses showed that en31FnCas9 exhibited greater fluctuations in the HNH and REC1 domains, with critical residues playing a significant role in mismatch recognition. MM/GBSA calculations supported these findings, revealing that en31FnCas9 had the highest binding affinity for sgRNA-tDNA complexes, mediated predominantly by the REC3 domain, with reduced contributions from the HNH and RuvC domains. This domain-specific interaction pattern highlights the unique mechanism by which en31FnCas9 achieves its superior specificity. The en31FnCas9 variant, with its triple mutation (G1243T/E1369R/E1449H) and REC2 deletion, demonstrates remarkable performance in mismatch discrimination, targeting a broader range of SNPs, and achieving higher cleavage efficiency. These findings emphasize the role of non-canonical base-pairing, domain flexibility, and enhanced REC3-mediated interactions in driving the superior specificity of en31FnCas9. This work not only advances the understanding of FnCas9 specificity but also lays the groundwork for designing more efficient and precise genome editing tools. Future studies can build upon these insights to further optimize Cas9 systems for clinical and research applications.

Objective 4: Expanding FnCas9 Functionality: Comparative Analysis of RNA and DNA Targeting

The final objective focused on understanding the structural and binding dynamics of FnCas9 when interacting with target RNA (tRNA) and tDNA substrates, aiming to uncover the mechanistic differences between the two. Specifically, the study examines how tRNA interacts with FnCas9 compared to tDNA, identifies the key residues involved in substrate recognition, and determines whether these residues overlap or differ between tRNA and tDNA binding. Additionally, the role of specific domains, such as the WED-PI domain, in modulating binding and cleavage efficiency is assessed. The work also explores the conformational changes and structural dynamics induced by tRNA binding, with the aim of understanding how these differences impact the functional properties of FnCas9. By addressing these questions, the study seeks to provide insights that could inform the design of FnCas9 variants with enhanced specificity and activity for RNA-targeting applications.

Through molecular dynamics simulations, contact analysis, binding energy calculations (MM/GBSA), and principal component analysis (PCA), this study reveals distinct interaction networks for tRNA and tDNA substrates. In tDNA-bound systems, residues primarily from the REC1 and REC3 domains (e.g., 106, 450, 451, 452, 453, 458, 663, 789, 794, 798, 809, 810, 845, 848, 856, and 1042) are predominantly involved in substrate recognition. Conversely, in tRNA-bound systems, contacts shift to include residues from the REC1, REC3, WED domains, and the L2 loop, with fewer contacts observed in the RuvC domain, suggesting a reduced involvement of this domain in tRNA binding. Unique residues—857, 858, and 1082–1084—were identified as specific to tRNA recognition, highlighting differences in the structural interactions required for RNA binding. Interestingly, eight residues (450, 451, 452, 453, 789, 809, 810, and 663) are common to both tRNA and tDNA, underscoring shared recognition elements while emphasizing distinct interaction patterns between the substrates. The results indicate that while tDNA exhibits a stronger overall binding affinity to FnCas9 compared to tRNA, as shown by MM/GBSA analysis, the WED and PI domains

show a preference for tRNA binding. This preference suggests that tRNA binding induces conformational changes that modulate binding affinity across different regions of the protein. PCA analysis further reveals differences in flexibility and conformational dynamics between the two substrates. Specifically, the RuvC domain exhibits increased fluctuations in tRNA-bound systems, confirming greater structural variability during RNA binding, which may influence cleavage activity. RMSF analysis supports these findings, with higher fluctuations observed in tRNA-bound systems for residues 857, 858, 1082, and 1083, while tDNA-bound systems exhibit more stability in the RuvC domain. Distance-based analysis also demonstrates that tRNA binding induces greater dynamic conformational shifts compared to tDNA, particularly in the WED and RuvC domains. The findings highlight the unique interaction networks and domain dynamics involved in RNA binding, with tRNA inducing greater flexibility and structural variability in FnCas9 compared to tDNA. These insights not only deepen our understanding of the molecular mechanisms underlying RNA and DNA targeting by FnCas9 but also provide a foundation for designing optimized Cas9 variants for RNA-targeting applications with enhanced specificity and activity.

These objectives highlight the critical role of structural dynamics and domain interactions in determining the specificity, efficiency, and versatility of Cas9 systems. From the fundamental differences in the gRNA-binding-induced domain rearrangements in SpCas9 and FnCas9 to the enhanced specificity seen in FnCas9 and its engineered variants, each part of this study emphasizes the importance of Cas9's flexible and dynamic nature. The understanding that domain rearrangements, particularly those in the REC3 and HNH domains, contribute significantly to the cleavage mechanism and specificity in FnCas9 builds on earlier insights and sets the stage for future optimization strategies. Furthermore, the comparative analysis of tRNA and tDNA binding in FnCas9 opens new directions for extending the CRISPR/Cas9 technology to RNA editing applications.

7.0.2 Future Work

Future research should focus on further elucidating the mechanisms behind Cas9 specificity and efficiency by examining the dynamics of DNA-bound ternary complexes and utilizing quantum mechanical simulations to explore the bond formation and breaking during cleavage. The study of engineered high-fidelity Cas9 variants and their allosteric modulation through targeted mutations or deletions could offer new avenues for enhancing Cas9's precision. Additionally, understanding the structural factors that govern FnCas9's interactions with different substrates, including tRNA, could inform the development of Cas9 variants optimized for RNA targeting. In vitro cleavage assays are strongly recommended to assess the fidelity of Cas9 variants on both on-target and predicted off-target DNA substrates Anderson *et al.* (2015); Tsai *et al.* (2015); Kim *et al.* (2015). Site-directed mutagenesis of residues identified in per-residue energy decomposition and allosteric analyses should be followed by cleavage profiling to determine their role in specificity Chen *et al.* (2017*b*). Additionally, biophysical assays such asSurface Plasmon Resonance (SPR) or Isothermal Titration Calorimetry (ITC) can measure binding affinities of wild-type and off-target Cas9 proteins to different DNA targets, validating MM/GBSA-based binding estimates. For RNA targeting

(Chapter 6), in vitro RNA cleavage assays with FnCas9 are essential to verify residues implicated in tRNA binding and processing Samai and et al. (2015); Smargon and et al. (2017). These experimental validations will complement our computational findings. Through these approaches, we can continue to refine Cas9-based genome editing technologies, improving their efficiency, precision, and applicability across a wide range of genetic engineering tasks.

REFERENCES

1. **Abadi, S.**, **W. X. Yan**, **D. Amar**, and **I. Mayrose** (2017). A machine learning approach for predicting crispr/cas9 off-target effects. *Nature Communications*, **8**, 1–10.

2. **Abraham, M. J.**, **T. Murtola**, **R. Schulz**, **S. Páll**, **J. C. Smith**, **B. Hess**, and **E. Lindahl** (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1-2**, 19–25. ISSN 2352-7110. URL https://www.sciencedirect.com/science/article/pii/S2352711015000059.

3. **Acharya, S.**, **A. H. Ansari**, **P. Kumar Das**, **S. Hirano**, **M. Aich**, **R. Rauthan**, **S. Mahato**, **S. Maddileti**, **S. Sarkar**, **M. Kumar**, *et al.* (2024). Pam-flexible engineered fncas9 variants for robust and ultra-precise genome editing and diagnostics. *Nature Communications*, **15**(1), 5471.

4. **Acharya, S.**, **A. Mishra**, **D. Paul**, **A. H. Ansari**, **M. Azhar**, **M. Kumar**, **R. Rauthan**, **N. Sharma**, **M. Aich**, **D. Sinha**, **S. Sharma**, **S. Jain**, **A. Ray**, **S. Jain**, **S. Ramalingam**, **S. Maiti**, and **D. Chakraborty** (2019*a*). Francisella novicida Cas9 interrogates genomic DNA with very high specificity and can be used for mammalian genome editing. *Proceedings of the National Academy of Sciences*, **116**(42), 20959–20968. ISSN 0027-8424. URL https://www.pnas.org/content/116/42/20959.

5. **Acharya, S.**, **A. Mishra**, **D. Paul**, **A. H. Ansari**, **M. Azhar**, **M. Kumar**, **R. Rauthan**, **N. Sharma**, **M. Aich**, **D. Sinha**, *et al.* (2019*b*). Francisella novicida cas9 interrogates genomic dna with very high specificity and can be used for mammalian genome editing. *Proceedings of the National Academy of Sciences*, **116**(42), 20959–20968.

6. **Agudelo, D.**, **S. Carter**, **M. Velimirovic**, **A. Duringer**, **J.-F. Rivest**, **S. Levesque**, **J. Loehr**, **M. Mouchiroud**, **D. Cyr**, **P. J. Waters**, *et al.* (2020). Versatile and robust genome editing with streptococcus thermophilus crispr1-cas9. *Genome research*, **30**(1), 107–117.

7. **Ahmad, I.** (2022). Crispr/cas9—a promising therapeutic tool to cure blindness: Current scenario and future prospects. *International Journal of Molecular Sciences*, **23**(19), 11482.

8. **Allen, M. P.** and **D. J. Tildesley** (2017). Computer simulation of liquids. *Oxford University Press*.

9. **Anders, C.**, **O. Niewoehner**, **A. Duerst**, and **M. Jinek** (2014*a*). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, **513**(7519), 569–573. ISSN 1476-4687. URL https://doi.org/10.1038/nature13579.

10. **Anders, C.**, **O. Niewoehner**, **A. Duerst**, and **M. Jinek** (2014*b*). Structural basis of pam-dependent target dna recognition by the cas9 endonuclease. *Nature*, **513**(7519), 569–573.

11. **Anderson, E. M.**, **A. Haupt**, and **et al.** (2015). Targeting dna with a minimal crispr-cas system. *Science*, **351**(6276), 1458–1461.

12. **Anderson, K. R.**, **M. Haeussler**, **C. Watanabe**, **V. Janakiraman**, **J. Lund**, **Z. Modrusan**, **J. Stinson**, **Q. Bei**, **A. Buechler**, **C. Yu**, **S. R. Thamminana**, **L. Tam**, **M.-A. Sowick**, **T. Alcantar**, **N. O'Neil**, **J. Li**, **L. Ta**, **L. Lima**, **M. Roose-Girma**, **X. Rairdan**, **S. Durinck**, and **S. Warming** (2018). CRISPR off-target analysis in genetically engineered rats and mice. *Nature Methods*, **15**(7), 512–514. ISSN 1548-7105. URL `https://doi.org/10.1038/s41592-018-0011-5`.

13. **Antoniou, P.**, **A. Miccio**, and **M. Brusson** (2021). Base and prime editing technologies for blood disorders. *Frontiers in Genome Editing*, **3**, 1. ISSN 2673-3439. URL `https://www.frontiersin.org/article/10.3389/fgeed.2021.618406`.

14. **Aqvist, J.** (1990). Ion-water interaction potentials derived from free energy perturbation simulations. *The Journal of Physical Chemistry*, **94**(21), 8021–8024.

15. **Åqvist, J.**, **V. B. Luzhkov**, and **B. O. Brandsdal** (2002). Ligand binding affinities from md simulations. *Accounts of chemical research*, **35**(6), 358–365.

16. **Aregger, M.**, **K. Xing**, and **T. Gonatopoulos-Pournatzis** (2021). Application of chymera cas9-cas12a combinatorial genome-editing platform for genetic interaction mapping and gene fragment deletion screening. *Nature protocols*, **16**(10), 4722–4765.

17. **Arya, M.**, **H. Sastry G**, **A. Motwani**, **S. Kumar**, and **A. Zaguia** (2022). A novel extra tree ensemble optimized dl framework (eteodl) for early detection of diabetes. *Frontiers in Public Health*, **9**, 797877.

18. **Ashkenazy, H.**, **S. Abadi**, **E. Martz**, **O. Chay**, **I. Mayrose**, **T. Pupko**, and **N. Ben-Tal** (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, **44**(W1), W344–W350. ISSN 0305-1048. URL `https://doi.org/10.1093/nar/gkw408`.

19. **Babcock, M. S.**, **E. P. Pednault**, and **W. K. Olson** (1994). Nucleic acid structure analysis: mathematics for local cartesian and helical structure parameters that are truly comparable between structures. *Journal of molecular biology*, **237**(1), 125–156.

20. **Bae, S.**, **J. Park**, and **J.-S. Kim** (2014). Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics (Oxford, England)*, **30**(10), 1473–1475. ISSN 1367-4811 (Electronic).

21. **Baek, M.** and **et al.** (2021). Accuracy of protein structure prediction in casp14. *Proteins: Structure, Function, and Bioinformatics*, **89**, 1052–1067.

22. **Baker, N. A.**, **D. Sept**, **S. Joseph**, **M. J. Holst**, and **J. A. McCammon** (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(18), 10037–10041. ISSN 0027-8424 (Print).

23. **Banás, P.**, **D. Hollas**, **M. Zgarbová**, **P. Jurecka**, **M. Orozco**, **T. E. Cheatham III**, **J. Sponer**, and **M. Otyepka** (2010). Performance of molecular mechanics force fields for rna simulations: stability of uucg and gnra hairpins. *Journal of Chemical Theory and Computation*, **6**(12), 3836–3849.

24. **Barrangou, R.**, **C. Fremaux**, **H. Deveau**, **M. Richards**, **P. Boyaval**, **S. Moineau**, **D. A. Romero**, and **P. Horvath** (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, **315**(5819), 1709–1712. ISSN 1095-9203 (Electronic).

25. **Barth, E.** and **T. Schlick** (1998). Overcoming stability limitations in biomolecular dynamics. i. combining force splitting via extrapolation with langevin dynamics or stochastic dynamics. *The Journal of Chemical Physics*, **109**(5), 1617–1632.

26. **Beeman, D.** (1976). Some multistep methods for use in molecular dynamics calculations. *Journal of Computational Physics*, **20**(2), 130–139.

27. **Bekker, H.** and **et al.** (2004). Molecular dynamics simulations: Advances and applications. *Advances in Applied Mathematics*, **32**, 1–20.

28. **Berendsen, H. J.** (1999). Molecular dynamics simulations: The limits and beyond. *Computational Molecular Dynamics: Challenges, Methods, Ideas*, 3–36.

29. **Berendsen, H. J.**, **J. P. Postma**, **W. F. van Gunsteren**, **A. DiNola**, and **J. R. Haak** (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, **81**(8), 3684–3690.

30. **Bergonzo, C.** and **T. E. Cheatham III** (2015). Multidimensional replica exchange molecular dynamics yields a converged ensemble of an rna tetranucleotide. *Journal of chemical theory and computation*, **11**(2), 396–406.

31. **Berman, H.**, **J. Westbrook**, **Z. Feng**, **G. Gilliland**, **T. Bhat**, **H. Weissig**, **I. Shindyalov**, and **P. Bourne** (2000). The protein data bank. *Nucleic Acids Research*, **28**, 235–242. URL `https://doi.org/10.1093/nar/28.1.235`.

32. **Bernardi, R. C.**, **M. C. R. Melo**, and **K. Schulten** (2014). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Annual Review of Biophysics*, **43**, 1–20.

33. **Best, R. B.**, **X. Zhu**, **J. Shim**, **P. E. Lopes**, **J. Mittal**, **M. Feig**, and **A. D. MacKerell Jr** (2012). Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation*, **8**(9), 3257–3273.

34. **Bhaya, D.**, **M. Davison**, and **R. Barrangou** (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*, **45**, 273–297. ISSN 1545-2948 (Electronic).

35. **Blaese, R. M.**, **K. W. Culver**, **A. D. Miller**, **C. S. Carter**, **T. Fleisher**, **M. Clerici**, **G. Shearer**, **L. Chang**, **Y. Chiang**, **P. Tolstoshev**, *et al.* (1995). T lymphocyte-directed gene therapy for ada-scid: initial trial results after 4 years. *Science*, **270**(5235), 475–480.

36. **Bland, C.**, **T. L. Ramsey**, **F. Sabree**, **M. Lowe**, **K. Brown**, **N. C. Kyrpides**, and **P. Hugenholtz** (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**(1), 209. ISSN 1471-2105. URL https://doi.org/10.1186/1471-2105-8-209.

37. **Boyle, E. A.**, **J. O. Andreasson**, **L. M. Chircus**, **S. H. Sternberg**, **M. J. Wu**, **C. K. Guegler**, **J. A. Doudna**, and **W. J. Greenleaf** (2017). High-throughput biochemical profiling reveals sequence determinants of dcas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences*, **114**(21), 5461–5466.

38. **Boyle, E. A.**, **W. R. Becker**, **H. B. Bai**, **J. S. Chen**, **J. A. Doudna**, and **W. J. Greenleaf** (2021). Quantification of cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Science Advances*, **7**(8), eabe5496.

39. **Braff, J. L.**, **S. J. Yaung**, **K. M. Esvelt**, and **G. M. Church** (2016). Characterization of Cas9–guide RNA orthologs. *Cold Spring Harbor Protocols*, **2016**(5), 422–425. ISSN 15596095.

40. **Brandsdal, B. O.**, **F. Österberg**, **M. Almlöf**, **I. Feierberg**, **V. B. Luzhkov**, and **J. Åqvist** (2003). Free energy calculations and ligand binding. *Advances in protein chemistry*, **66**, 123–158.

41. **Bravo, J. P.**, **M.-S. Liu**, **G. N. Hibshman**, **T. L. Dangerfield**, **K. Jung**, **R. S. McCool**, **K. A. Johnson**, and **D. W. Taylor** (2022*a*). Structural basis for mismatch surveillance by crispr–cas9. *Nature*, **603**(7900), 343–347.

42. **Bravo, J. P. K.**, **M.-S. Liu**, **G. N. Hibshman**, **T. L. Dangerfield**, **K. Jung**, **R. S. McCool**, **K. A. Johnson**, and **D. W. Taylor** (2022*b*). Structural basis for mismatch surveillance by CRISPR–Cas9. *Nature*, **603**(7900), 343–347. ISSN 1476-4687. URL https://doi.org/10.1038/s41586-022-04470-1.

43. **Brouns, S. J.**, **M. M. Jore**, **M. Lundgren**, **E. R. Westra**, **R. J. Slijkhuis**, **A. P. Snijders**, **M. J. Dickman**, **K. S. Makarova**, **E. V. Koonin**, and **J. Van Der Oost** (2008*a*). Small crispr rnas guide antiviral defense in prokaryotes. *Science*, **321**(5891), 960–964.

44. **Brouns, S. J. J.**, **M. M. Jore**, **M. Lundgren**, **E. R. Westra**, **R. J. H. Slijkhuis**, **A. P. L. Snijders**, **M. J. Dickman**, **K. S. Makarova**, **E. V. Koonin**, and **J. van der Oost** (2008*b*). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science (New York, N.Y.)*, **321**(5891), 960–964. ISSN 1095-9203 (Electronic).

45. **Brown, S. e. a.** (2024). Exploring rna-based drug targets in oncology. *Expert Opinion on Drug Discovery*, **19**(3), 101–117. URL https://www.tandfonline.com/doi/full/10.1080/17460441.2024.2313455.

46. **Bussi, G.**, **D. Donadio**, and **M. Parrinello** (2007). Canonical sampling through velocity rescaling. *The Journal of chemical physics*, **126**(1), 014101.

47. **Cameron, P.**, **C. Fuller**, **P. Donohoue**, **B. Jones**, **M. Thompson**, **M. Carter**, **S. Gradia**, **B. Vidal**, **E. Garner**, and **E. Slorach** (2017*a*). Mapping the genomic landscape of crispr–cas9 cleavage. *Nature Methods*, **14**, 600–606.

48. **Cameron, P.**, **C. K. Fuller**, **P. D. Donohoue**, **B. N. Jones**, **M. S. Thompson**, **M. M. Carter**, **S. Gradia**, **B. Vidal**, **E. Garner**, **E. M. Slorach**, *et al.* (2017*b*). Mapping the genomic landscape of crispr–cas9 cleavage. *Nature methods*, **14**(6), 600–606.

49. **Carpenter, J. C.** and **G. Lignani** (2021). Gene editing and modulation: the holy grail for the genetic epilepsies? *Neurotherapeutics*, **18**(3), 1515–1523.

50. **Casalino, L.**, **L. Nierzwicki**, **M. Jinek**, and **G. Palermo** (2020). Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR-Cas9 Revealed by Ab Initio Molecular Dynamics. *ACS Catalysis*, **10**(22), 13596–13605. URL https://doi.org/10.1021/acscatal.0c03566.

51. **Case, D. A.**, **H. M. Aktulga**, **K. Belfon**, **D. S. Cerutti**, **G. A. Cisneros**, **V. W. D. Cruzeiro**, **N. Forouzesh**, **T. J. Giese**, **A. W. Götz**, **H. Gohlke**, **S. Izadi**, **K. Kasavajhala**, **M. C. Kaymak**, **E. King**, **T. Kurtzman**, **T.-S. Lee**, **P. Li**, **J. Liu**, **T. Luchko**, **R. Luo**, **M. Manathunga**, **M. R. Machado**, **H. M. Nguyen**, **K. A. O'Hearn**, **A. V. Onufriev**, **F. Pan**, **S. Pantano**, **R. Qi**, **A. Rahnamoun**, **A. Risheh**, **S. Schott-Verdugo**, **A. Shajan**, **J. Swails**, **J. Wang**, **H. Wei**, **X. Wu**, **Y. Wu**, **S. Zhang**, **S. Zhao**, **Q. Zhu**, **T. E. I. I. I. Cheatham**, **D. R. Roe**, **A. Roitberg**, **C. Simmerling**, **D. M. York**, **M. C. Nagan**, and **K. M. J. Merz** (2023). AmberTools. *Journal of Chemical Information and Modeling*, **63**(20), 6183–6191. ISSN 1549-9596. URL https://doi.org/10.1021/acs.jcim.3c01153.

52. **Casini, A.**, **M. Olivieri**, **G. Petris**, **C. Montagna**, **G. Reginato**, **G. Maule**, **F. Lorenzin**, **D. Prandi**, **A. Romanel**, **F. Demichelis**, **A. Inga**, and **A. Cereseto** (2018*a*). A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nature Biotechnology*, **36**(3), 265–271. ISSN 1546-1696. URL https://doi.org/10.1038/nbt.4066.

53. **Casini, A.**, **M. Olivieri**, **G. Petris**, **C. Montagna**, **G. Reginato**, **G. Maule**, **F. Lorenzin**, **D. Prandi**, **A. Romanel**, **F. Demichelis**, *et al.* (2018*b*). A highly specific spcas9 variant is identified by in vivo screening in yeast. *Nature biotechnology*, **36**(3), 265–271.

54. **Charlesworth, C. T.**, **P. S. Deshpande**, **D. P. Dever**, **J. Camarena**, **V. T. Lemgart**, **M. K. Cromer**, **C. A. Vakulskas**, **M. A. Collingwood**, **L. Zhang**, **N. M. Bode**, **M. A. Behlke**, **B. Dejene**, **B. Cieniewicz**, **R. Romano**, **B. J. Lesch**, **N. Gomez-Ospina**, **S. Mantri**, **M. Pavel-Dinu**, **K. I. Weinberg**, and **M. H. Porteus** (2019*a*). Identification of preexisting adaptive immunity to Cas9 proteins in humans. *Nature medicine*, **25**(2), 249–254. ISSN 1546-170X (Electronic).

55. **Charlesworth, C. T.**, **P. S. Deshpande**, **D. P. Dever**, **J. Camarena**, **V. T. Lemgart**, **M. K. Cromer**, **C. A. Vakulskas**, **M. A. Collingwood**, **L. Zhang**, **N. M. Bode**, *et al.* (2019*b*). Iden-

tification of preexisting adaptive immunity to cas9 proteins in humans. *Nature medicine*, **25**(2), 249–254.

56. **Chen, F.**, **M. Alphonse**, and **Q. Liu** (2020*a*). Strategies for nonviral nanoparticle-based delivery of crispr/cas9 therapeutics. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, **12**(3), e1609.

57. **Chen, H.**, **J. Choi**, and **S. Bailey** (2014). Cut Site Selection by the Two Nuclease Domains of the Cas9 RNA-guided Endonuclease *. *Journal of Biological Chemistry*, **289**(19), 13284–13294. ISSN 0021-9258. URL https://doi.org/10.1074/jbc.M113.539726.

58. **Chen, I.-M. A.**, **K. Chu**, **K. Palaniappan**, **M. Pillay**, **A. Ratner**, **J. Huang**, **M. Huntemann**, **N. Varghese**, **J. R. White**, **R. Seshadri**, **T. Smirnova**, **E. Kirton**, **S. P. Jungbluth**, **T. Woyke**, **E. A. Eloe-Fadrosh**, **N. N. Ivanova**, and **N. C. Kyrpides** (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic acids research*, **47**(D1), D666–D677. ISSN 1362-4962 (Electronic).

59. **Chen, J. S.**, **Y. S. Dagdas**, **B. P. Kleinstiver**, **M. M. Welch**, **A. A. Sousa**, **L. B. Harrington**, **S. H. Sternberg**, **J. K. Joung**, **A. Yildiz**, and **J. A. Doudna** (2017*a*). Enhanced proofreading governs crispr–cas9 targeting accuracy. *Nature*, **550**(7676), 407–410.

60. **Chen, J. S.**, **Y. S. Dagdas**, **B. P. Kleinstiver**, **M. M. Welch**, **A. A. Sousa**, **L. B. Harrington**, **S. H. Sternberg**, **J. K. Joung**, **A. Yildiz**, and **J. A. Doudna** (2017*b*). Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature*, **550**(7676), 407–410. ISSN 1476-4687. URL https://doi.org/10.1038/nature24268.

61. **Chen, K.**, **L. A. Kurgan**, and **J. Ruan** (2007). Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Structural Biology*, **7**(1), 25. ISSN 1472-6807. URL https://doi.org/10.1186/1472-6807-7-25.

62. **Chen, Q.**, **Y. Zhang**, and **H. Yin** (2021). Recent advances in chemical modifications of guide RNA, mRNA and donor template for CRISPR-mediated genome editing. *Advanced Drug Delivery Reviews*, **168**, 246–258. ISSN 0169-409X. URL https://www.sciencedirect.com/science/article/pii/S0169409X20301599.

63. **Chen, S.**, **Y. Yao**, **Y. Zhang**, and **G. Fan** (2020*b*). Crispr system: Discovery, development and off-target detection. *Cellular signalling*, **70**, 109577.

64. **Chen, S.-J.** (2019). Minimizing off-target effects in CRISPR-Cas9 genome editing.

65. **Cho, S. W.**, **S. Kim**, **Y. Kim**, **J. Kweon**, **H. S. Kim**, **S. Bae**, and **J.-S. Kim** (2014*a*). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome research*, **24**(1), 132–141. ISSN 1549-5469 (Electronic).

66. **Cho, S. W.**, **S. Kim**, **Y. Kim**, **J. Kweon**, **H. S. Kim**, **S. Bae**, and **J.-S. Kim** (2014*b*). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome research*, **24**(1), 132–141. ISSN 1549-5469 (Electronic).

67. **Chuai, G.**, **H. Ma**, **J. Yan**, **M. Chen**, **N. Hong**, **D. Xue**, **C. Zhou**, **C. Zhu**, **K. Chen**, and **B. Duan** (2018). Deepcrispr: Optimized crispr guide rna design by deep learning. *Genome Biology*, **19**, 80.

68. **Chylinski, K.**, **A. Le Rhun**, and **E. Charpentier** (2013). The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA biology*, **10**(5), 726–737. ISSN 1555-8584 (Electronic).

69. **Chylinski, K.**, **K. S. Makarova**, **E. Charpentier**, and **E. V. Koonin** (2014). Classification and evolution of type II CRISPR-Cas systems. *Nucleic acids research*, **42**(10), 6091–6105. ISSN 1362-4962 (Electronic).

70. **Clementel, D.**, **A. Del Conte**, **A. M. Monzon**, **G. F. Camagni**, **G. Minervini**, **D. Piovesan**, and **S. C. Tosatto** (2022). Ring 3.0: Fast generation of probabilistic residue interaction networks from structural ensembles. *Nucleic Acids Research*, **50**(W1), W651–W656.

71. **Colque-Navarro, P.**, **G. Jacobsson**, **R. Andersson**, **J.-I. Flock**, and **R. Mollby** (2010). Levels of antibody against 11 staphylococcus aureus antigens in a healthy population. *Clinical and Vaccine Immunology*, **17**(7), 1117–1123.

72. **Cong, L.**, **F. A. Ran**, **D. Cox**, **S. Lin**, **R. Barretto**, **N. Habib**, **P. D. Hsu**, **X. Wu**, **W. Jiang**, **L. A. Marraffini**, and **F. Zhang** (2013*a*). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, **339**(6121), 819–823. ISSN 1095-9203 (Electronic).

73. **Cong, L.**, **F. A. Ran**, **D. Cox**, **S. Lin**, **R. Barretto**, **N. Habib**, **P. D. Hsu**, **X. Wu**, **W. Jiang**, **L. A. Marraffini**, *et al.* (2013*b*). Multiplex genome engineering using crispr/cas systems. *Science*, **339**(6121), 819–823.

74. **Consortium, T. U.** (2016). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **45**(D1), D158–D169. ISSN 0305-1048. URL `https://doi.org/10.1093/nar/gkw1099`.

75. **Cornell, W. D.**, **P. Cieplak**, **C. I. Bayly**, **I. R. Gould**, **K. M. Merz**, **D. M. Ferguson**, **D. C. Spellmeyer**, **T. Fox**, **J. W. Caldwell**, and **P. A. Kollman** (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, **117**(19), 5179–5197.

76. **Costa, J. R.**, **B. E. Bejcek**, **J. E. McGee**, **A. I. Fogel**, **K. R. Brimacombe**, and **R. Ketteler** (2017). Genome editing using engineered nucleases and their use in genomic screening. *Assay Guidance Manual [Internet]*.

77. **Crosetto, N.**, **A. Mitra**, **M. Silva**, **M. Bienko**, **N. Dojer**, **Q. Wang**, **E. Karaca**, **R. Chiarle**, **M. Skrzypczak**, and **K. Ginalski** (2013). Nucleotide-resolution dna double-strand break mapping by next-generation sequencing. *Nature Methods*, **10**, 361–365.

78. **Dagdas, Y. S.**, **J. S. Chen**, **S. H. Sternberg**, **J. A. Doudna**, and **A. Yildiz** (2017). A conformational checkpoint between dna binding and cleavage by crispr-cas9. *Science advances*, **3**(8), eaao0027.

79. **Darden, T.**, **D. York**, and **L. Pedersen** (1993). Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, **98**(12), 10089–10092. ISSN 0021-9606. URL https://doi.org/10.1063/1.464397.

80. **Darve, E.**, **M. Rodriguez**, and **A. Pohorille** (2008). Adaptive biasing force method for efficient sampling of complex systems. *Physical Review E*, **77**, 066701.

81. **Delhove, J.**, **I. Osenk**, **I. Prichard**, and **M. Donnelley** (2020). Public Acceptability of Gene Therapy and Gene Editing for Human Use: A Systematic Review. *Human Gene Therapy*, **31**(1-2), 20–46. URL https://doi.org/10.1089/hum.2019.197.

82. **Deltcheva, E.**, **K. Chylinski**, **C. M. Sharma**, **K. Gonzales**, **Y. Chao**, **Z. A. Pirzada**, **M. R. Eckert**, **J. Vogel**, and **E. Charpentier** (2011). Crispr rna maturation by trans-encoded small rna and host factor rnase iii. *Nature*, **471**(7340), 602–607.

83. **Deuflhard, P.** and **M. Weber** (2005). Robust perron cluster analysis in conformation dynamics. *Linear algebra and its applications*, **398**, 161–184.

84. **DeWitt, M. A.**, **W. Magis**, **N. L. Bray**, **T. Wang**, **J. R. Berman**, **F. Urbinati**, **S.-J. Heo**, **T. Mitros**, **D. P. Muñoz**, **D. Boffelli**, **D. B. Kohn**, **M. C. Walters**, **D. Carroll**, **D. I. K. Martin**, and **J. E. Corn** (2016). Selection-free genome editing of the sickle mutation in human adult hematopoietic stem/progenitor cells. *Science translational medicine*, **8**(360), 360ra134. ISSN 1946-6242 (Electronic).

85. **Dhainaut, M.**, **S. A. Rose**, **G. Akturk**, **A. Wroblewska**, **S. R. Nielsen**, **E. S. Park**, **M. Buckup**, **V. Roudko**, **L. Pia**, **R. Sweeney**, *et al.* (2022). Spatial crispr genomics identifies regulators of the tumor microenvironment. *Cell*, **185**(7), 1223–1239.

86. **Doench, J.**, **E. Hartenian**, **D. Graham**, **Z. Tothova**, **M. Hegde**, **I. Smith**, **M. Sullender**, **B. Ebert**, **R. Xavier**, and **D. Root** (2014). Rational design of highly active sgrnas for crispr-cas9–mediated gene inactivation. *Nature Biotechnology*, **32**, 1262–1267. CrossRef.

87. **Doench, J. G.**, **N. Fusi**, **M. Sullender**, **M. Hegde**, **E. W. Vaimberg**, **K. F. Donovan**, **I. Smith**, **Z. Tothova**, **C. Wilen**, **R. Orchard**, *et al.* (2016). Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, **34**(2), 184–191.

88. **Dolinsky, T. J.**, **J. E. Nielsen**, **J. A. McCammon**, and **N. A. Baker** (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic acids research*, **32**(Web Server issue), W665–7. ISSN 1362-4962 (Electronic).

89. **Donohoue, P. D.**, **M. Pacesa**, **E. Lau**, **B. Vidal**, **M. J. Irby**, **D. B. Nyer**, **T. Rotstein**, **L. Banh**, **M. S. Toh**, **J. Gibson**, **B. Kohrs**, **K. Baek**, **A. L. G. Owen**, **E. M. Slorach**, **M. van Overbeek**, **C. K. Fuller**, **A. P. May**, **M. Jinek**, and **P. Cameron** (2021). Conformational control of Cas9 by CRISPR hybrid RNA-DNA guides mitigates off-target activity in T cells. *Molecular cell*, **81**(17), 3637–3649.e5. ISSN 1097-4164 (Electronic).

90. **Doudna, J.** and **E. Charpentier** (2014). Genome editing. the new frontier of genome engineering with crispr-cas9.(2014). *Science*, **346**, 6213.

91. **Duan, J.**, **G. Lu**, **Z. Xie**, **M. Lou**, **J. Luo**, **L. Guo**, and **Y. Zhang** (2014). Genome-wide identification of crispr/cas9 off-targets in human genome. *Cell research*, **24**(8), 1009–1012.

92. **East, K. W.**, **J. C. Newton**, **U. N. Morzan**, **Y. B. Narkhede**, **A. Acharya**, **E. Skeens**, **G. Jogl**, **V. S. Batista**, **G. Palermo**, and **G. P. Lisi** (2020). Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *Journal of the American Chemical Society*, **142**(3), 1348–1358. ISSN 15205126.

93. **Edgar, R. C.** (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC bioinformatics*, **8**, 18. ISSN 1471-2105 (Electronic).

94. **Estarellas, C.**, **M. Otyepka**, **J. Koča**, **P. Banáš**, **M. Krepl**, and **J. Šponer** (2015). Molecular dynamic simulations of protein/RNA complexes: CRISPR/Csy4 endoribonuclease. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1850**(5), 1072–1090. ISSN 0304-4165. URL `https://www.sciencedirect.com/science/article/pii/S0304416514003572`.

95. **Ewaisha, R.** and **K. S. Anderson** (2023). Immunogenicity of crispr therapeutics—critical considerations for clinical translation. *Frontiers in Bioengineering and Biotechnology*, **11**, 1138596.

96. **Ewald, P. P.** (1921). Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, **64**, 253–287.

97. **Feng, S.**, **Z. Wang**, **A. Li**, **X. Xie**, **J. Liu**, **S. Li**, **Y. Li**, **B. Wang**, **L. Hu**, **L. Yang**, and **T. Guo** (2021). Strategies for High-Efficiency Mutation Using the CRISPR/Cas System. *Frontiers in cell and developmental biology*, **9**, 803252. ISSN 2296-634X (Print).

98. **Fonfara, I.**, **A. Le Rhun**, **K. Chylinski**, **K. S. Makarova**, **A.-L. Lécrivain**, **J. Bzdrenga**, **E. V. Koonin**, and **E. Charpentier** (2014). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic acids research*, **42**(4), 2577–2590. ISSN 1362-4962 (Electronic).

99. **Frangoul, H.**, **D. Altshuler**, **M. D. Cappellini**, **Y.-S. Chen**, **J. Domm**, **B. K. Eustace**, **J. Foell**, **J. de la Fuente**, **S. Grupp**, **R. Handgretinger**, *et al.* (2021). Crispr-cas9 gene editing for sickle cell disease and β-thalassemia. *New England Journal of Medicine*, **384**(3), 252–260.

100. **Fu, Y.**, **J. A. Foden**, **C. Khayter**, **M. L. Maeder**, **D. Reyon**, **J. K. Joung**, and **J. D. Sander** (2013). High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells. *Nature biotechnology*, **31**(9), 822–826.

101. **Fu, Y.**, **J. D. Sander**, **D. Reyon**, **V. M. Cascio**, and **J. K. Joung** (2014*a*). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature Biotechnology*, **32**(3), 279–284. ISSN 1546-1696. URL `https://doi.org/10.1038/nbt.2808`.

102. **Fu, Y.**, **J. D. Sander**, **D. Reyon**, **V. M. Cascio**, and **J. K. Joung** (2014*b*). Improving crispr-cas nuclease specificity using truncated guide rnas. *Nature biotechnology*, **32**(3), 279–284.

103. **Fuchsbauer, O.**, **P. Swuec**, **C. Zimberger**, **B. Amigues**, **S. Levesque**, **D. Agudelo**, **A. Duringer**, **A. Chaves-Sanjuan**, **S. Spinelli**, **G. M. Rousseau**, **M. Velimirovic**, **M. Bolognesi**, **A. Roussel**, **C. Cambillau**, **S. Moineau**, **Y. Doyon**, and **A. Goulet** (2019). Cas9 Allosteric Inhibition by the Anti-CRISPR Protein AcrIIA6. *Molecular Cell*, **76**(6), 922–937.e7. ISSN 10974164. URL https://www.cell.com/molecular-cell/pdf/S1097-2765(19)30697-5.pdf.

104. **Gasiunas, G.**, **J. K. Young**, **T. Karvelis**, **D. Kazlauskas**, **T. Urbaitis**, **M. Jasnauskaite**, **M. M. Grusyte**, **S. Paulraj**, **P.-H. Wang**, **Z. Hou**, **S. K. Dooley**, **M. Cigan**, **C. Alarcon**, **N. D. Chilcoat**, **G. Bigelyte**, **J. L. Curcuru**, **M. Mabuchi**, **Z. Sun**, **R. T. Fuchs**, **E. Schildkraut**, **P. R. Weigele**, **W. E. Jack**, **G. B. Robb**, **Č. Venclovas**, and **V. Siksnys** (2020). A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nature Communications*, **11**(1), 5512. ISSN 2041-1723. URL https://doi.org/10.1038/s41467-020-19344-1.

105. **Genheden, S.** and **U. Ryde** (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert opinion on drug discovery*, **10**(5), 449–461. ISSN 1746-045X (Electronic).

106. **Georgiadis, C.**, **J. Rasaiyaah**, **S. A. Gkazi**, **R. Preece**, **A. Etuk**, **A. Christi**, and **W. Qasim** (2021). Base-edited car t cells for combinational therapy against t cell malignancies. *Leukemia*, **35**(12), 3466–3481.

107. **Ghosh, D.**, **P. Venkataramani**, **S. Nandi**, and **S. Bhattacharjee** (2019). CRISPR–Cas9 a boon or bane: the bumpy road ahead to cancer therapeutics. *Cancer Cell International*, **19**(1), 12. ISSN 1475-2867. URL https://doi.org/10.1186/s12935-019-0726-0.

108. **Ghouneimy, A.**, **A. Mahas**, **T. Marsic**, **R. Aman**, and **M. Mahfouz** (2022). Crispr-based diagnostics: challenges and potential solutions toward point-of-care applications. *ACS Synthetic Biology*, **12**(1), 1–16.

109. **Gonatopoulos-Pournatzis, T.**, **M. Aregger**, **K. R. Brown**, **S. Farhangmehr**, **U. Braunschweig**, **H. N. Ward**, **K. C. Ha**, **A. Weiss**, **M. Billmann**, **T. Durbic**, *et al.* (2020). Genetic interaction mapping and exon-resolution functional genomics with a hybrid cas9–cas12a platform. *Nature biotechnology*, **38**(5), 638–648.

110. **Gramegna, A.** and **P. Giudici** (2022). Shapley Feature Selection. *FinTech*, **1**(1), 72–80. ISSN 2674-1032.

111. **Gupta, S.**, **A. Singh**, *et al.* (2023). Recent advances in crispr/cas9 technology for genetic engineering. *Nucleic Acids Research*, **51**(3), 245–265. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10138764/.

112. **Guy, A.**, **T. Piggot**, and **S. Khalid** (2012). Single-Stranded DNA within Nanopores: Conformational Dynamics and Implications for Sequencing; a Molecular Dynamics Simulation Study. *Biophysical Journal*, **103**(5), 1028–1036. ISSN 0006-3495. URL https://doi.org/10.1016/j.bpj.2012.08.012.

113. **Guyon, I.**, **J. Weston**, **S. Barnhill**, and **V. Vapnik** (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, **46**, 389–422.

114. **Haft, D. H.**, **J. Selengut**, **E. F. Mongodin**, and **K. E. Nelson** (2005). A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLOS Computational Biology*, **1**(6), 1–10. URL https://doi.org/10.1371/journal.pcbi.0010060.

115. **Hagberg, A.**, **P. Swart**, and **D. S Chult** (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

116. **Hamelberg, D.**, **J. Mongan**, and **J. A. McCammon** (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics*, **120**(24), 11919–11929.

117. **Hanlon, K.**, **B. Kleinstiver**, **S. Garcia**, **M. Zaborowski**, **A. Volak**, **S. Spirig**, **A. Muller**, **A. Sousa**, **S. Tsai**, and **N. Bengtsson** (2019). High levels of aav vector integration into crispr-induced dna breaks. *Nature Communications*, **10**, 1–11.

118. **Hatano, T.** and **S.-i. Sasa** (2001). Steady-State Thermodynamics of Langevin Systems. *Phys. Rev. Lett.*, **86**(16), 3463–3466. URL https://link.aps.org/doi/10.1103/PhysRevLett.86.3463.

119. **Hess, B.** (2008). P-lincs: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation*, **4**(1), 116–122.

120. **Hirano, H.**, **J. S. Gootenberg**, **T. Horii**, **O. O. Abudayyeh**, **M. Kimura**, **P. D. Hsu**, **T. Nakane**, **R. Ishitani**, **I. Hatada**, **F. Zhang**, **H. Nishimasu**, and **O. Nureki** (2016). Structure and Engineering of Francisella novicida Cas9. *Cell*, **164**(5), 950–961. ISSN 1097-4172 (Electronic).

121. **Hirano, S.**, **O. O. Abudayyeh**, **J. S. Gootenberg**, **T. Horii**, **R. Ishitani**, **I. Hatada**, **F. Zhang**, **H. Nishimasu**, and **O. Nureki** (2019). Structural basis for the promiscuous PAM recognition by Corynebacterium diphtheriae Cas9. *Nature Communications*, **10**(1), 1968. ISSN 2041-1723. URL https://doi.org/10.1038/s41467-019-09741-6.

122. **HMADA, C.**, **K. Belfon**, **I. Ben-Shalom**, **J. Berryman**, **S. Brozell**, **D. Cerutti**, **T. Cheatham**, **G. Cisneros**, **V. Cruzeiro**, **T. Darden**, *et al.* (2022). Amber2022. *San Francisco.: University of California*.

123. **Hockney, R. W.** (1970). The potential calculation and some applications. *Methods in Computational Physics*, **9**, 135–211.

124. **Hoover, W. G.** (1985). Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, **31**(3), 1695–1697.

125. **Hospital, A.** and **et al.** (2015). Molecular dynamics simulations: Advances and applications. *Molecular Systems Biology*, **11**, 1–14.

126. **Howe, S. J.**, **M. R. Mansour**, **K. Schwarzwaelder**, **C. Bartholomae**, **M. Hubank**, **H. Kempski**, **M. H. Brugman**, **K. Pike-Overzet**, **S. J. Chatters**, **D. De Ridder**, *et al.* (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of scid-x1 patients. *The Journal of clinical investigation*, **118**(9).

127. **Hsu, P. D.**, **D. A. Scott**, **J. A. Weinstein**, **F. A. Ran**, **S. Konermann**, **V. Agarwala**, **Y. Li**, **E. J. Fine**, **X. Wu**, **O. Shalem**, *et al.* (2013). Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, **31**(9), 827–832.

128. **Hu, J. H.**, **S. M. Miller**, **M. H. Geurts**, **W. Tang**, **L. Chen**, **N. Sun**, **C. M. Zeina**, **X. Gao**, **H. A. Rees**, **Z. Lin**, *et al.* (2018). Evolved cas9 variants with broad pam compatibility and high dna specificity. *Nature*, **556**(7699), 57–63.

129. **Huai, C.**, **G. Li**, **R. Yao**, **Y. Zhang**, **M. Cao**, **L. Kong**, **C. Jia**, **H. Yuan**, **H. Chen**, **D. Lu**, and **Q. Huang** (2017). Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nature communications*, **8**(1), 1375. ISSN 2041-1723 (Electronic).

130. **Hui, S.** (2008). Overview of drug delivery and alternative methods to electroporation. *Methods in Molecular Biology*, **423**, 91–107.

131. **Humphrey, W.**, **A. Dalke**, and **K. Schulten** (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, **14**(1), 27–28,33–38. ISSN 0263-7855 (Print).

132. **Hustedt, N.** and **D. Durocher** (2017). The control of dna repair by the cell cycle. *Nature cell biology*, **19**(1), 1–9.

133. **Ihry, R. J.**, **K. A. Worringer**, **M. R. Salick**, **E. Frias**, **D. Ho**, **K. Theriault**, **S. Kommineni**, **J. Chen**, **M. Sondey**, **C. Ye**, *et al.* (2018). p53 inhibits crispr–cas9 engineering in human pluripotent stem cells. *Nature medicine*, **24**(7), 939–946.

134. **Ivanov, I. E.**, **A. V. Wright**, **J. C. Cofsky**, **K. D. P. Aris**, **J. A. Doudna**, and **Z. Bryant** (2020). Cas9 interrogates dna in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences*, **117**(11), 5853–5860.

135. **Jacquin, A.**, **D. Odom**, and **M. Lukk** (2019). Crisflash: Open-source software to generate crispr guide rnas against genomes annotated with individual variation. *Bioinformatics*, **35**, 3146–3147.

136. **Jansen, R.**, **J. D. v. Embden**, **W. Gaastra**, and **L. M. Schouls** (2002). Identification of genes that are associated with dna repeats in prokaryotes. *Molecular microbiology*, **43**(6), 1565–1575.

137. **Jiang, F.** and **J. A. Doudna** (2015). The structural biology of crispr-cas systems. *Current opinion in structural biology*, **30**, 100–111.

138. **Jiang, F.** and **J. A. Doudna** (2017). CRISPR–Cas9 Structures and Mechanisms. *Annual Review of Biophysics*, **46**(1), 505–529. URL https://doi.org/10.1146/annurev-biophys-062215-010822.

139. **Jiang, F.**, **D. W. Taylor**, **J. S. Chen**, **J. E. Kornfeld**, **K. Zhou**, **A. J. Thompson**, **E. Nogales**, and **J. A. Doudna** (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science (New York, N.Y.)*, **351**(6275), 867–871. ISSN 1095-9203 (Electronic).

140. **Jiang, F.**, **K. Zhou**, **L. Ma**, **S. Gressel**, and **J. A. Doudna** (2015). STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science (New York, N.Y.)*, **348**(6242), 1477–1481. ISSN 1095-9203 (Electronic).

141. **Jinek, M.**, **K. Chylinski**, **I. Fonfara**, **M. Hauer**, **J. A. Doudna**, and **E. Charpentier** (2012*a*). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, **337**(6096), 816–821. ISSN 1095-9203 (Electronic).

142. **Jinek, M.**, **K. Chylinski**, **I. Fonfara**, **M. Hauer**, **J. A. Doudna**, and **E. Charpentier** (2012*b*). A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, **337**(6096), 816–821.

143. **Jinek, M.**, **F. Jiang**, **D. W. Taylor**, **S. H. Sternberg**, **E. Kaya**, **E. Ma**, **C. Anders**, **M. Hauer**, **K. Zhou**, **S. Lin**, **M. Kaplan**, **A. T. Iavarone**, **E. Charpentier**, **E. Nogales**, and **J. A. Doudna** (2014*a*). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science (New York, N.Y.)*, **343**(6176), 1247997. ISSN 1095-9203 (Electronic).

144. **Jinek, M.**, **F. Jiang**, **D. W. Taylor**, **S. H. Sternberg**, **E. Kaya**, **E. Ma**, **C. Anders**, **M. Hauer**, **K. Zhou**, **S. Lin**, *et al.* (2014*b*). Structures of cas9 endonucleases reveal rna-mediated conformational activation. *Science*, **343**(6176), 1247997.

145. **Jones, S. K.**, **J. A. Hawkins**, **N. V. Johnson**, **C. Jung**, **K. Hu**, **J. R. Rybarski**, **J. S. Chen**, **J. A. Doudna**, **W. H. Press**, and **I. J. Finkelstein** (2021). Massively parallel kinetic profiling of natural and engineered crispr nucleases. *Biophysical Journal*, **120**(3), 138a.

146. **Jorgensen, W. L.**, **J. Chandrasekhar**, **J. D. Madura**, **R. W. Impey**, and **M. L. Klein** (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, **79**(2), 926–935. ISSN 0021-9606. URL https://doi.org/10.1063/1.445869.

147. **Josephs, E. A.**, **D. D. Kocak**, **C. J. Fitzgibbon**, **J. McMenemy**, **C. A. Gersbach**, and **P. E. Marszalek** (2015*a*). Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic acids research*, **43**(18), 8924–8941. ISSN 1362-4962 (Electronic).

148. **Josephs, E. A.**, **D. D. Kocak**, **C. J. Fitzgibbon**, **J. McMenemy**, **C. A. Gersbach**, and **P. E. Marszalek** (2015*b*). Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic acids research*, **43**(18), 8924–8941. ISSN 1362-4962 (Electronic).

149. **Joung, J.** and **S. Tsai** (2015). Genomewide unbiased identification of dsbs evaluated by sequencing (guide-seq). *Nature Biotechnology*, **33**, 187–197.

150. **Jurrus, E.**, **D. Engel**, **K. Star**, **K. Monson**, **J. Brandi**, **L. E. Felberg**, **D. H. Brookes**, **L. Wilson**, **J. Chen**, **K. Liles**, **M. Chun**, **P. Li**, **D. W. Gohara**, **T. Dolinsky**, **R. Konecny**, **D. R. Koes**, **J. E. Nielsen**, **T. Head-Gordon**, **W. Geng**, **R. Krasny**, **G.-W. Wei**, **M. J. Holst**, **J. A. McCammon**, and **N. A. Baker** (2018). Improvements to the APBS biomolecular solvation software suite. *Protein Science*, **27**(1), 112–128. ISSN 0961-8368. URL `https://doi.org/10.1002/pro.3280`.

151. **Karvelis, T.**, **J. K. Young**, and **V. Siksnys** (2019). A pipeline for characterization of novel Cas9 orthologs. *Methods in enzymology*, **616**, 219–240. ISSN 1557-7988 (Electronic).

152. **Kim, D.**, **S. Bae**, **J. Park**, **E. Kim**, **S. Kim**, **H. Yu**, **J. Hwang**, **J. Kim**, and **J. Kim** (2015). Digenome-seq: Genome-wide profiling of crispr-cas9 off-target effects in human cells. *Nature Methods*, **12**, 237–243.

153. **Kim, H.** and **J.-S. Kim** (2014). A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics*, **15**(5), 321–334.

154. **Kim, S.**, **T. Bae**, **J. Hwang**, and **J.-S. Kim** (2017). Rescue of high-specificity Cas9 variants using sgRNAs with matched 5' nucleotides. *Genome biology*, **18**(1), 218. ISSN 1474-760X (Electronic).

155. **Kim, Y.**, **S. J. Lee**, **H. J. Yoon**, **N. K. Kim**, **B. J. Lee**, and **J. Y. Suh** (2019). Anti-CRISPR AcrIIC3 discriminates between Cas9 orthologs via targeting the variable surface of the HNH nuclease domain. *FEBS Journal*, **286**(23), 4661–4674. ISSN 17424658.

156. **Klein, M. L.** and **et al.** (2018). Gaussian accelerated molecular dynamics: A new method for accelerating molecular dynamics simulations. *Journal of Chemical Theory and Computation*, **14**, 3094–3104.

157. **Klein, M. L.** and **et al.** (2022). Machine learning in crispr/cas9 applications: A review. *Nature Reviews Genetics*, **23**, 1–15.

158. **Kleinstiver, B. P.**, **V. Pattanayak**, **M. S. Prew**, **S. Q. Tsai**, **N. T. Nguyen**, **Z. Zheng**, and **J. K. Joung** (2016*a*). High-fidelity crispr–cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**(7587), 490–495.

159. **Kleinstiver, B. P.**, **V. Pattanayak**, **M. S. Prew**, **S. Q. Tsai**, **N. T. Nguyen**, **Z. Zheng**, and **J. K. Joung** (2016*b*). High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**(7587), 490–495. ISSN 1476-4687. URL `https://doi.org/10.1038/nature16526`.

160. **Konermann, S.**, **M. D. Brigham**, **A. E. Trevino**, **J. Joung**, **O. O. Abudayyeh**, **C. Barcena**, **P. D. Hsu**, **N. Habib**, **J. S. Gootenberg**, **H. Nishimasu**, **O. Nureki**, and **F. Zhang** (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**(7536), 583–588. ISSN 1476-4687 (Electronic).

161. **Koo, T.**, **J. Lee**, and **J.-S. Kim** (2015). Measuring and reducing off-target activities of programmable nucleases including crispr-cas9. *Molecules and cells*, **38**(6), 475.

162. **Koonin, E. V.** and **K. S. Makarova** (2019). Origins and evolution of CRISPR-Cas systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **374**(1772), 20180087. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2018.0087.

163. **Koonin, E. V.**, **K. S. Makarova**, and **F. Zhang** (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Current opinion in microbiology*, **37**, 67–78. ISSN 1879-0364 (Electronic).

164. **Kostyushev, D.**, **S. Brezgin**, **A. Kostyusheva**, **D. Zarifyan**, **I. Goptar**, and **V. Chulanov** (2019). Orthologous crispr/cas9 systems for specific and efficient degradation of covalently closed circular dna of hepatitis b virus. *Cellular and Molecular Life Sciences*, **76**(9), 1779–1794.

165. **Kožić, M.** and **B. Bertoša** (2024). Trajectory maps: molecular dynamics visualization and analysis. *NAR Genomics and Bioinformatics*, **6**(1), lqad114.

166. **Kramer, O.** and **O. Kramer** (2016). Scikit-learn. *Machine learning for evolution strategies*, 45–53.

167. **Kräutler, V.**, **W. F. Van Gunsteren**, and **P. H. Hünenberger** (2001). A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, **22**(5), 501–508.

168. **Kursa, M. B.** and **W. R. Rudnicki** (2010). Feature selection with the boruta package. *Journal of statistical software*, **36**, 1–13.

169. **Labun, K.**, **T. Montague**, **J. Gagnon**, **S. Thyme**, and **E. Valen** (2016). Chopchop v2: A web tool for the next generation of crispr genome engineering. *Nucleic Acids Research*, **44**, W272–W276.

170. **Lapinaite, A.**, **G. J. Knott**, **C. M. Palumbo**, **E. Lin-Shiao**, **M. F. Richter**, **K. T. Zhao**, **P. A. Beal**, **D. R. Liu**, and **J. A. Doudna** (2020). Dna capture by a crispr-cas9–guided adenine base editor. *Science*, **369**(6503), 566–571. ISSN 0036-8075. URL https://science.sciencemag.org/content/369/6503/566.

171. **Leach, A. R.**, *Molecular Modelling: Principles and Applications*. Prentice Hall, Harlow, UK, 2001.

172. **Lee, C. M.**, **T. J. Cradick**, and **G. Bao** (2016). The neisseria meningitidis crispr-cas9 system enables specific genome editing in mammalian cells. *Molecular Therapy*, **24**(3), 645–654.

173. **Lee, J. K.**, **E. Jeong**, **J. Lee**, **M. Jung**, **E. Shin**, **Y.-h. Kim**, **K. Lee**, **I. Jung**, **D. Kim**, **S. Kim**, and **J.-S. Kim** (2018*a*). Directed evolution of CRISPR-Cas9 to increase its specificity. *Nature Communications*, **9**(1), 3048. ISSN 2041-1723. URL https://doi.org/10.1038/s41467-018-05477-x.

174. **Lee, J. K.**, **E. Jeong**, **J. Lee**, **M. Jung**, **E. Shin**, **Y.-h. Kim**, **K. Lee**, **I. Jung**, **D. Kim**, **S. Kim**, *et al.* (2018*b*). Directed evolution of crispr-cas9 to increase its specificity. *Nature communications*, **9**(1), 3048.

175. **Lessard, S.**, **L. Francioli**, **J. Alfoldi**, **J.-C. Tardif**, **P. T. Ellinor**, **D. G. MacArthur**, **G. Lettre**, **S. H. Orkin**, and **M. C. Canver** (2017). Human genetic variation alters crispr-cas9 on-and off-targeting specificity at therapeutically implicated loci. *Proceedings of the National Academy of Sciences*, **114**(52), E11257–E11266.

176. **Li, H.**, **Y. Yang**, **W. Hong**, **M. Huang**, **M. Wu**, and **X. Zhao** (2020). Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduction and Targeted Therapy*, **5**(1), 1. ISSN 2059-3635. URL `https://doi.org/10.1038/s41392-019-0089-y`.

177. **Li, J.**, **S. Hong**, **W. Chen**, **E. Zuo**, and **H. Yang** (2019). Advances in detecting and reducing off-target effects generated by crispr-mediated genome editing. *Journal of genetics and genomics*, **46**(11), 513–521.

178. **Lin, Y.**, **T. Cradick**, **M. Brown**, **H. Deshmukh**, **P. Ranjan**, **N. Sarode**, **B. Wile**, **P. Vertino**, **F. Stewart**, and **G. Bao** (2014). Crispr/cas9 systems have off-target activity with insertions or deletions between target dna and guide rna sequences. *Nucleic Acids Research*, **42**, 7473–7485.

179. **Lindorff-Larsen, K.**, **S. Piana**, **R. O. Dror**, and **D. E. Shaw** (2012). How fast-folding proteins fold. *Science*, **334**(6055), 517–520.

180. **Lino, C.**, **J. Harper**, **J. Carney**, and **J. Timlin** (2018). Delivering crispr: A review of the challenges and approaches. *Drug Delivery*, **25**, 1234–1257.

181. **Listgarten, J.** *et al.* (2018). Machine-learning-based crispr guide design. *Broad Institute*. URL `https://www.broadinstitute.org/what-broad/areas-focus/project-spotlight/questions-and-answers-about-crispr`.

182. **Liu, Q.**, **F. Yang**, **J. Zhang**, **H. Liu**, **S. Rahman**, **S. Islam**, **W. Ma**, and **M. She** (2021). Application of crispr/cas9 in crop quality improvement. *International Journal of Molecular Sciences*, **22**(8), 4206.

183. **Liu, Q.**, **H. Zhang**, and **X. Huang** (2020). Anti-CRISPR proteins targeting the CRISPR-Cas system enrich the toolkit for genetic engineering. *The FEBS journal*, **287**(4), 626–644. ISSN 1742-4658 (Electronic).

184. **Lu, X.-J.** and **W. K. Olson** (2003). 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic acids research*, **31**(17), 5108–5121.

185. **Lu, Y.**, **J. Xue**, **T. Deng**, **X. Zhou**, **K. Yu**, **L. Deng**, **M. Huang**, **X. Yi**, **M. Liang**, **Y. Wang**, *et al.* (2020). Safety and feasibility of crispr-edited t cells in patients with refractory non-small-cell lung cancer. *Nature medicine*, **26**(5), 732–740.

186. **Lundberg, S. M.** and **S.-I. Lee** (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, **30**.

187. **Maier, J. A.**, **C. Martinez**, **K. Kasavajhala**, **L. Wickstrom**, **K. E. Hauser**, and **C. Simmerling** (2015). ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, **11**(8), 3696–3713.

188. **Maji, B.**, **S. A. Gangopadhyay**, **M. Lee**, **M. Shi**, **P. Wu**, **R. Heler**, **B. Mok**, **D. Lim**, **S. U. Siriwardena**, **B. Paul**, **V. Dančík**, **A. Vetere**, **M. F. Mesleh**, **L. A. Marraffini**, **D. R. Liu**, **P. A. Clemons**, **B. K. Wagner**, and **A. Choudhary** (2019). A High-Throughput Platform to Identify Small-Molecule Inhibitors of CRISPR-Cas9. *Cell*, **177**(4), 1067–1079.e19. ISSN 1097-4172 (Electronic).

189. **Makarova, K. S.**, **D. H. Haft**, **R. Barrangou**, **S. J. J. Brouns**, **E. Charpentier**, **P. Horvath**, **S. Moineau**, **F. J. M. Mojica**, **Y. I. Wolf**, **A. F. Yakunin**, **J. van der Oost**, and **E. V. Koonin** (2011). Evolution and classification of the CRISPR-Cas systems.

190. **Makarova, K. S.** and **E. V. Koonin** (2015). Annotation and Classification of CRISPR-Cas Systems. *Methods in molecular biology (Clifton, N.J.)*, **1311**, 47–75. ISSN 1940-6029 (Electronic).

191. **Makarova, K. S.**, **Y. I. Wolf**, **O. S. Alkhnbashi**, **F. Costa**, **S. A. Shah**, **S. J. Saunders**, **R. Barrangou**, **S. J. J. Brouns**, **E. Charpentier**, **D. H. Haft**, **P. Horvath**, **S. Moineau**, **F. J. M. Mojica**, **R. M. Terns**, **M. P. Terns**, **M. F. White**, **A. F. Yakunin**, **R. A. Garrett**, **J. van der Oost**, **R. Backofen**, and **E. V. Koonin** (2015). An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology*, **13**(11), 722–736. ISSN 1740-1534. URL `https://doi.org/10.1038/nrmicro3569`.

192. **Makarova, K. S.**, **Y. I. Wolf**, **J. Iranzo**, **S. A. Shmakov**, **O. S. Alkhnbashi**, **S. J. J. Brouns**, **E. Charpentier**, **D. Cheng**, **D. H. Haft**, **P. Horvath**, **S. Moineau**, **F. J. M. Mojica**, **D. Scott**, **S. A. Shah**, **V. Siksnys**, **M. P. Terns**, **Č. Venclovas**, **M. F. White**, **A. F. Yakunin**, **W. Yan**, **F. Zhang**, **R. A. Garrett**, **R. Backofen**, **J. van der Oost**, **R. Barrangou**, and **E. V. Koonin** (2020). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology*, **18**(2), 67–83. ISSN 1740-1534. URL `https://doi.org/10.1038/s41579-019-0299-x`.

193. **Mali, P.**, **L. Yang**, **K. M. Esvelt**, **J. Aach**, **M. Guell**, **J. E. DiCarlo**, **J. E. Norville**, and **G. M. Church** (2013). RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, **339**(6121), 823–826. ISSN 1095-9203 (Electronic).

194. **Manghwar, H.**, **B. Li**, **X. Ding**, **A. Hussain**, **K. Lindsey**, **X. Zhang**, and **S. Jin** (2020). Crispr/cas systems in genome editing: methodologies and tools for sgrna design, off-target evaluation, and strategies to mitigate off-target effects. *Advanced science*, **7**(6), 1902312.

195. **Manghwar, H.**, **K. Lindsey**, **X. Zhang**, and **S. Jin** (2019). CRISPR/Cas System: Recent Advances and Future Prospects for Genome Editing. *Trends in Plant Science*, **24**(12),

1102–1125. ISSN 1360-1385. URL https://www.sciencedirect.com/science/article/pii/S1360138519302432.

196. **Marino, N. D.**, **J. Y. Zhang**, **A. L. Borges**, **A. A. Sousa**, **M. Lina**, **B. J. Rauch**, **R. T. Walton**, **J. D. Berry**, **J. K. Joung**, **B. P. Kleinstiver**, and **J. Bondy-denomy** (2019). Discovery of widespread type i and type v crispr-cas inhibitors. *Science*, **362**(6411), 240–242.

197. **Martinez Velazquez, L. A.** and **B. G. Ballios** (2021). The next generation of molecular and cellular therapeutics for inherited retinal disease. *International Journal of Molecular Sciences*, **22**(21), 11542.

198. **Martyna, G. J.**, **D. J. Tobias**, and **M. L. Klein** (1994). Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, **101**(5), 4177–4189.

199. **McGibbon, R. T.**, **K. A. Beauchamp**, **M. P. Harrigan**, **C. Klein**, **J. M. Swails**, **C. X. Hernández**, **C. R. Schwantes**, **L.-P. Wang**, **T. J. Lane**, and **V. S. Pande** (2015). Mdtraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal*, **109**(8), 1528–1532.

200. **McKenna, A.** and **J. Shendure** (2018). Flashfry: A fast and flexible tool for large-scale crispr target design. *BMC Biology*, **16**, 74.

201. **Meaker, G. A.**, **E. J. Hair**, and **T. E. Gorochowski** (2020*a*). Advances in engineering CRISPR-Cas9 as a molecular Swiss Army knife. *Synthetic Biology*, **5**(1). ISSN 2397-7000. URL https://doi.org/10.1093/synbio/ysaa021.

202. **Meaker, G. A.**, **E. J. Hair**, and **T. E. Gorochowski** (2020*b*). Advances in engineering CRISPR-Cas9 as a molecular swiss army knife. *Synthetic Biology*, **5**(1). ISSN 23977000.

203. **Mehta, A.** and **O. M. Merkel** (2020*a*). Immunogenicity of cas9 protein. *Journal of pharmaceutical sciences*, **109**(1), 62–67.

204. **Mehta, A.** and **O. M. Merkel** (2020*b*). Immunogenicity of Cas9 Protein. *Journal of pharmaceutical sciences*, **109**(1), 62–67. ISSN 1520-6017. URL https://pubmed.ncbi.nlm.nih.gov/31589876https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7115921/.

205. **Miao, Y.**, **V. A. Feher**, and **J. A. McCammon** (2015). Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *Journal of chemical theory and computation*, **11**(8), 3584–3595.

206. **Miao, Y.** and **J. A. McCammon** (2016). Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. *Molecular simulation*, **42**(13), 1046–1055.

207. **Mitchell, B. P.**, **R. V. Hsu**, **M. A. Medrano**, **N. T. Zewde**, **Y. B. Narkhede**, and **G. Palermo** (2020). Spontaneous embedding of dna mismatches within the rna: Dna hybrid of crispr-cas9. *Frontiers in Molecular Biosciences*, **7**, 39.

208. **Moen, M. T.** and **I. G. Johnston** (2023). Hyperhmm: efficient inference of evolutionary and progressive dynamics on hypercubic transition graphs. *Bioinformatics*, **39**(1), btac803.

209. **Mojica, F. J.**, **G. Juez**, and **F. Rodriguez-Valera** (1993). Transcription at different salinities of haloferax mediterranei sequences adjacent to partially modified psti sites. *Molecular microbiology*, **9**(3), 613–621.

210. **Moreno-Mateos, M. A.**, **C. E. Vejnar**, **J.-D. Beaudoin**, **J. P. Fernandez**, **E. K. Mis**, **M. K. Khokha**, and **A. J. Giraldez** (2015). Crisprscan: designing highly efficient sgrnas for crispr-cas9 targeting in vivo. *Nature methods*, **12**(10), 982–988.

211. **Motta, B. M.**, **P. P. Pramstaller**, **A. A. Hicks**, and **A. Rossini** (2017). The impact of crispr/cas9 technology on cardiac research: from disease modelling to therapeutic approaches. *Stem cells international*, **2017**(1), 8960236.

212. **Mout, R.**, **M. Ray**, **Y.-W. Lee**, **F. Scaletti**, and **V. M. Rotello** (2017). In vivo delivery of crispr/cas9 for therapeutic gene editing: progress and challenges. *Bioconjugate chemistry*, **28**(4), 880–884.

213. **Müller, M.**, **C. M. Lee**, **G. Gasiunas**, **T. H. Davis**, **T. J. Cradick**, **V. Siksnys**, **G. Bao**, **T. Cathomen**, and **C. Mussolino** (2016). Streptococcus thermophilus crispr-cas9 systems enable specific editing of the human genome. *Molecular Therapy*, **24**(3), 636–644.

214. **Nadakuduti, S. S.** and **F. Enciso-Rodríguez** (2021). Advances in Genome Editing With CRISPR Systems and Transformation Technologies for Plant DNA Manipulation. *Frontiers in Plant Science*, **11**, 2267. ISSN 1664-462X. URL `https://www.frontiersin.org/article/10.3389/fpls.2020.637159`.

215. **Naeem, M.**, **S. Majeed**, **M. Z. Hoque**, and **I. Ahmad** (2020). Latest developed strategies to minimize the off-target effects in crispr-cas-mediated genome editing. *Cells*, **9**(7), 1608.

216. **Naso, M.**, **B. Tomkowicz**, **W. Perry**, and **W. Strohl** (2017). Adeno-associated virus (aav) as a vector for gene therapy. *BioDrugs*, **31**, 317–334.

217. **Nishimasu, H.**, **L. Cong**, **W. X. Yan**, **F. A. Ran**, **B. Zetsche**, **Y. Li**, **A. Kurabayashi**, **R. Ishitani**, **F. Zhang**, and **O. Nureki** (2015). Crystal Structure of Staphylococcus aureus Cas9. *Cell*, **162**(5), 1113–1126. ISSN 1097-4172 (Electronic).

218. **Nishimasu, H.**, **F. A. Ran**, **P. D. Hsu**, **S. Konermann**, **S. I. Shehata**, **N. Dohmae**, **R. Ishitani**, **F. Zhang**, and **O. Nureki** (2014*a*). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**(5), 935–949. ISSN 1097-4172 (Electronic).

219. **Nishimasu, H.**, **F. A. Ran**, **P. D. Hsu**, **S. Konermann**, **S. I. Shehata**, **N. Dohmae**, **R. Ishitani**, **F. Zhang**, and **O. Nureki** (2014*b*). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**(5), 935–949. ISSN 1097-4172 (Electronic).

220. **Noé, F.**, **C. Schütte**, **E. Vanden-Eijnden**, **L. Reich**, and **T. R. Weikl** (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, **106**(45), 19011–19016.

221. **Nose, S.** and **M. L. Klein** (1986). Constant pressure molecular dynamics for molecular systems. *Physical Review B*, **33**(6), 339–342.

222. **Olson, W. K.**, **M. Bansal**, **S. K. Burley**, **R. E. Dickerson**, **M. Gerstein**, **S. C. Harvey**, **U. Heinemann**, **X.-J. Lu**, **S. Neidle**, **Z. Shakked**, *et al.* (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of molecular biology*, **313**(1), 229–237.

223. **Oostenbrink, C.**, **A. Villa**, **A. E. Mark**, and **W. F. Van Gunsteren** (2004*a*). A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force field. *Journal of Computational Chemistry*, **25**(13), 1656–1676.

224. **Oostenbrink, C.**, **A. Villa**, **A. E. Mark**, and **W. F. Van Gunsteren** (2004*b*). A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. *Journal of computational chemistry*, **25**(13), 1656–1676.

225. **Osuka, S.**, **K. Isomura**, **S. Kajimoto**, **T. Komori**, **H. Nishimasu**, **T. Shima**, **O. Nureki**, and **S. Uemura** (2018). Real-time observation of flexible domain movements in CRISPR-Cas9. *The EMBO journal*, **37**(10). ISSN 1460-2075 (Electronic).

226. **Pacesa, M.**, **C.-H. Lin**, **A. Cléry**, **A. Saha**, **P. R. Arantes**, **K. Bargsten**, **M. J. Irby**, **F. H.-T. Allain**, **G. Palermo**, **P. Cameron**, *et al.* (2022). Structural basis for cas9 off-target activity. *Cell*, **185**(22), 4067–4081.

227. **Palermo, G.** (2019*a*). Structure and Dynamics of the CRISPR–Cas9 Catalytic Complex. *Journal of Chemical Information and Modeling*, **59**(5), 2394–2406. ISSN 1549-9596. URL `https://doi.org/10.1021/acs.jcim.8b00988`.

228. **Palermo, G.** (2019*b*). Structure and Dynamics of the CRISPR–Cas9 Catalytic Complex. *Journal of Chemical Information and Modeling*, **59**(5), 2394–2406. ISSN 1549-9596. URL `https://doi.org/10.1021/acs.jcim.8b00988`.

229. **Palermo, G.**, **J. S. Chen**, **C. G. Ricci**, **I. Rivalta**, **M. Jinek**, **V. S. Batista**, **J. A. Doudna**, and **J. A. McCammon** (2018). Key role of the REC lobe during CRISPR-Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain. *Quarterly reviews of biophysics*, **51**. ISSN 1469-8994 (Electronic).

230. **Palermo, G.**, **Y. Miao**, **R. C. Walker**, **M. Jinek**, and **J. A. McCammon** (2016*a*). Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *ACS Central Science*, **2**(10), 756–763. ISSN 2374-7943. URL `https://doi.org/10.1021/acscentsci.6b00218`.

231. **Palermo, G.**, **Y. Miao**, **R. C. Walker**, **M. Jinek**, and **J. A. McCammon** (2016*b*). Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *ACS Central Science*, **2**(10), 756–763. ISSN 2374-7943. URL https://doi.org/10.1021/acscentsci.6b00218.

232. **Palermo, G.**, **Y. Miao**, **R. C. Walker**, **M. Jinek**, and **J. A. McCammon** (2017). CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proceedings of the National Academy of Sciences*, **114**(28), 7260–7265. URL https://www.pnas.org/doi/abs/10.1073/pnas.1707645114.

233. **Palermo, G.**, **C. G. Ricci**, and **J. A. McCammon** (2019). The invisible dance of CRISPR-Cas9. Simulations unveil the molecular side of the gene-editing revolution. *Physics today*, **72**(4), 30–36. ISSN 0031-9228 (Print).

234. **Panda, G.** and **A. Ray** (2022). Comparative structural and dynamics study of free and grna-bound fncas9 and spcas9 proteins. *Computational and Structural Biotechnology Journal*, **20**, 4172–4184.

235. **Park, P.** (2009). Chip–seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669–680. CrossRef.

236. **Pattanayak, V.**, **S. Lin**, **J. P. Guilinger**, **E. Ma**, **J. A. Doudna**, and **D. R. Liu** (2013*a*). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature Biotechnology*, **31**(9), 839–843. ISSN 1546-1696. URL https://doi.org/10.1038/nbt.2673.

237. **Pattanayak, V.**, **S. Lin**, **J. P. Guilinger**, **E. Ma**, **J. A. Doudna**, and **D. R. Liu** (2013*b*). High-throughput profiling of off-target dna cleavage reveals rna-programmed cas9 nuclease specificity. *Nature biotechnology*, **31**(9), 839–843.

238. **Pérez, A.**, **I. Marchán**, **D. Svozil**, **J. Sponer**, **T. E. r. Cheatham**, **C. A. Laughton**, and **M. Orozco** (2007). Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophysical journal*, **92**(11), 3817–3829. ISSN 0006-3495 (Print).

239. **Peterson, J.** *et al.* (2022). The promise of rna-targeted therapies in drug discovery. *Nature Reviews Drug Discovery*, **21**, 475–493. URL https://www.nature.com/articles/s41573-022-00521-4.

240. **Pettersen, E. F.**, **T. D. Goddard**, **C. C. Huang**, **G. S. Couch**, **D. M. Greenblatt**, **E. C. Meng**, and **T. E. Ferrin** (2004). Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, **25**(13), 1605–1612.

241. **Phillips, J. C.**, **R. Braun**, **W. Wang**, **J. Gumbart**, **E. Tajkhorshid**, **E. Villa**, **C. Chipot**, **R. D. Skeel**, **L. Kal**, and **K. Schulten** (2005). Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, **26**(16), 1781–1802.

242. **Piana, S.**, **K. Lindorff-Larsen**, and **D. E. Shaw** (2013). Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences*, **110**(15), 5915–5920.

243. **Ran, F. A.**, **P. D. Hsu**, **C.-Y. Lin**, **J. S. Gootenberg**, **S. Konermann**, **A. E. Trevino**, **D. A. Scott**, **A. Inoue**, **S. Matoba**, **Y. Zhang**, and **F. Zhang** (2013*a*). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**(6), 1380–1389. ISSN 1097-4172 (Electronic).

244. **Ran, F. A.**, **P. D. Hsu**, **J. Wright**, **V. Agarwala**, **D. A. Scott**, and **F. Zhang** (2013*b*). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, **8**(11), 2281–2308. ISSN 1750-2799. URL https://doi.org/10.1038/nprot.2013.143.

245. **Rauch, B. J.**, **M. R. Silvis**, **J. F. Hultquist**, **C. S. Waters**, **M. J. McGregor**, **N. J. Krogan**, and **J. Bondy-Denomy** (2017). Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell*, **168**(1-2), 150–158.e10. ISSN 1097-4172 (Electronic).

246. **Ray, A.** and **R. Di Felice** (2020). Protein-Mutation-Induced Conformational Changes of the DNA and Nuclease Domain in CRISPR/Cas9 Systems by Molecular Dynamics Simulations. *The journal of physical chemistry. B*, **124**(11), 2168–2179. ISSN 1520-5207 (Electronic).

247. **Rees, H. A.**, **C. Wilson**, **J. L. Doman**, and **D. R. Liu** (2019). Analysis and minimization of cellular rna editing by dna adenine base editors. *Science Advances*, **5**(5). URL https://advances.sciencemag.org/content/5/5/eaax5717.

248. **Ren, B.**, **F. Yan**, **Y. Kuang**, **N. Li**, **D. Zhang**, **H. Lin**, and **H. Zhou** (2017). A CRISPR/Cas9 toolkit for efficient targeted base editing to induce genetic variations in rice.

249. **Ren, X.**, **Z. Yang**, **J. Xu**, **J. Sun**, **D. Mao**, **Y. Hu**, **S. Yang**, **H. Qiao**, **X. Wang**, and **Q. Hu** (2014). Enhanced specificity and efficiency of the crispr/cas9 system with optimized sgrna parameters in drosophila. *Cell Reports*, **9**, 1151–1162. CrossRef, PubMed.

250. **Ricci, C. G.**, **J. S. Chen**, **Y. Miao**, **M. Jinek**, **J. A. Doudna**, **J. A. McCammon**, and **G. Palermo** (2019). Deciphering off-target effects in crispr-cas9 through accelerated molecular dynamics. *ACS central science*, **5**(4), 651–662.

251. **Rice, P.**, **I. Longden**, and **A. Bleasby** (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, **16**(6), 276–277. ISSN 0168-9525 (Print).

252. **Röblitz, S.** and **M. Weber** (2013). Fuzzy spectral clustering by pcca+: application to markov state models and data classification. *Advances in Data Analysis and Classification*, **7**, 147–179.

253. **Roe, D. R.** and **T. E. I. I. I. Cheatham** (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, **9**(7), 3084–3095. ISSN 1549-9618. URL https://doi.org/10.1021/ct400341p.

254. **Roe, D. R.** and **T. E. Cheatham III** (2013). Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, **9**(7), 3084–3095.

255. **Roshanravan, N.**, **H. Tutunchi**, **F. Najafipour**, **M. Dastouri**, **S. Ghaffari**, and **A. Jebeli** (2022). A glance at the application of crispr/cas9 gene-editing technology in cardiovascular diseases. *Journal of Cardiovascular and Thoracic Research*, **14**(2), 77.

256. **Ryan, D. E.**, **D. Taussig**, **I. Steinfeld**, **S. M. Phadnis**, **B. D. Lunstad**, **M. Singh**, **X. Vuong**, **K. D. Okochi**, **R. McCaffrey**, **M. Olesiak**, **S. Roy**, **C. W. Yung**, **B. Curry**, **J. R. Sampson**, **L. Bruhn**, and **D. J. Dellinger** (2018). Improving CRISPR-Cas specificity with chemical modifications in single-guide RNAs. *Nucleic acids research*, 792–803. ISSN 1362-4962. URL `https://pubmed.ncbi.nlm.nih.gov/29216382`.

257. **Ryckaert, J. P.**, **G. Ciccotti**, and **H. J. C. Berendsen** (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, **23**(3), 327–341.

258. **Safari, F.**, **G. Hatam**, **A. B. Behbahani**, **V. Rezaei**, **M. Barekati-Mowahed**, **P. Petramfar**, and **F. Khademi** (2020). Crispr system: a high-throughput toolbox for research and treatment of parkinson's disease. *Cellular and molecular neurobiology*, **40**, 477–493.

259. **Sakiyama, Y.** (1996). Gene therapy for adenosine deaminase deficiency. *[Hokkaido Igaku Zasshi] The Hokkaido Journal of Medical Science*, **71**(1), 27–32.

260. **Samai, P.** and **et al.** (2015). Co-transcriptional dna and rna cleavage during type iii crispr–cas immunity. *Cell*, **161**(5), 1164–1174.

261. **Scherer, M. K.**, **B. Trendelkamp-Schroer**, **F. Paul**, **G. Pérez-Hernández**, **M. Hoffmann**, **N. Plattner**, **C. Wehmeyer**, **J.-H. Prinz**, and **F. Noé** (2015). Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of chemical theory and computation*, **11**(11), 5525–5542.

262. **Schlick, T.**, **E. Barth**, and **M. Mandziuk** (1997). Biomolecular dynamics at long timesteps: bridging the timescale gap between simulation and experimentation. *Annual Review of Biophysics and Biomolecular Structure*, **26**, 181–222.

263. **Schmid-Burgk, J. L.**, **L. Gao**, **D. Li**, **Z. Gardner**, **J. Strecker**, **B. Lash**, and **F. Zhang** (2020). Highly Parallel Profiling of Cas9 Variant Specificity. *Molecular Cell*, **78**(4), 794–800.e8. ISSN 1097-2765. URL `https://www.sciencedirect.com/science/article/pii/S109727652030143X`.

264. **Schrodinger, L.** and **W. DeLano** (2020*a*). Pymol. URL `http://www.pymol.org/pymol`.

265. **Schrodinger, L.** and **W. DeLano** (2020*b*). PyMOL.

266. **Schwede, T.**, **J. Kopp**, **N. Guex**, and **M. C. Peitsch** (2003). Swiss-model: an automated protein homology-modeling server. *Nucleic acids research*, **31**(13), 3381–3385.

267. **Schütt, K. T.** and **et al.** (2017). Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, **8**, 13890.

268. **Sherkatghanad, Z.**, **M. Abdar**, **J. Charlier**, and **V. Makarenkov** (2023). Using traditional machine learning and deep learning methods for on- and off-target prediction in crispr/cas9: a review. *Briefings in Bioinformatics*, **24**(3).

269. **Shutaywi, M.** and **N. N. Kachouie** (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, **23**(6), 759.

270. **Singh, A.**, **D. Chakraborty**, and **S. Maiti** (2016). CRISPR/Cas9: a historical and chemical biology perspective of targeted genome engineering. *Chem. Soc. Rev.*, **45**(24), 6666–6684. URL `http://dx.doi.org/10.1039/C6CS00197A`.

271. **Slaymaker, I. M.**, **L. Gao**, **B. Zetsche**, **D. A. Scott**, **W. X. Yan**, and **F. Zhang** (2016*a*). Rationally engineered Cas9 nucleases with improved specificity. *Science (New York, N.Y.)*, **351**(6268), 84–88. ISSN 1095-9203 (Electronic).

272. **Slaymaker, I. M.**, **L. Gao**, **B. Zetsche**, **D. A. Scott**, **W. X. Yan**, and **F. Zhang** (2016*b*). Rationally engineered cas9 nucleases with improved specificity. *Science*, **351**(6268), 84–88.

273. **Smargon, A. A.** and **et al.** (2017). Cas13b is a type vi-b crispr-associated rna-guided rnase differentially regulated by accessory proteins csx27 and csx28. *Molecular Cell*, **65**(4), 618–630.

274. **Sternberg, S. H.**, **B. LaFrance**, **M. Kaplan**, and **J. A. Doudna** (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature*, **527**(7576), 110–113. ISSN 1476-4687 (Electronic).

275. **Sternberg, S. H.**, **S. Redding**, **M. Jinek**, **E. C. Greene**, and **J. A. Doudna** (2014). Dna interrogation by the crispr rna-guided endonuclease cas9. *Biophysical Journal*, **106**(2), 695a.

276. **Stone, D.**, **K. R. Long**, **M. A. Loprieno**, **H. S. D. S. Feelixge**, **E. J. Kenkel**, **R. M. Liley**, **S. Rapp**, **P. Roychoudhury**, **T. Nguyen**, **L. Stensland**, *et al.* (2021). Crispr-cas9 gene editing of hepatitis b virus in chronically infected humanized mice. *Molecular Therapy-Methods & Clinical Development*, **20**, 258–275.

277. **Sugano, S.**, **R. Nishihama**, **M. Shirakawa**, **J. Takagi**, **Y. Matsuda**, **S. Ishida**, **T. Shimada**, **I. Hara-Nishimura**, **K. Osakabe**, and **T. Kohchi** (2018). Efficient crispr/cas9-based genome editing and its application to conditional genetic analysis in marchantia polymorpha. *PLoS ONE*, **13**, e0205117.

278. **Sugita, Y.** and **Y. Okamoto** (2000). Replica exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, **329**, 261–270.

279. **Swope, W. C.**, **H. C. Andersen**, **P. H. Berens**, and **K. R. Wilson** (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, **76**(1), 637–649.

280. **Tangprasertchai, N. S.**, **R. Di Felice**, **X. Zhang**, **I. M. Slaymaker**, **C. Vazquez Reyes**, **W. Jiang**, **R. Rohs**, and **P. Z. Qin** (2017). CRISPR-Cas9 Mediated DNA Unwinding Detected Using Site-Directed Spin Labeling. *ACS chemical biology*, **12**(6), 1489–1493. ISSN 1554-8937 (Electronic).

281. **Team, A.** (2021). Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.

282. **Torrie, G. M.** and **J. P. Valleau** (1977). Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, **23**, 187–199.

283. **Trozzi, F.**, **X. Wang**, and **P. Tao** (2021). Umap as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: a comparison study. *The Journal of Physical Chemistry B*, **125**(19), 5022–5034.

284. **Tsai, S. Q.**, **Z. Zheng**, **N. T. Nguyen**, **M. Liebers**, **V. V. Topkar**, **V. Thapar**, **N. Wyvekens**, **C. Khayter**, **A. J. Iafrate**, **L. P. Le**, **M. J. Aryee**, and **J. K. Joung** (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, **33**(2), 187–197. ISSN 1546-1696. URL `https://doi.org/10.1038/nbt.3117`.

285. **Turq, P.**, **F. Lantelme**, and **H. L. Friedman** (1977). Brownian dynamics: Its application to ionic solutions. *The Journal of Chemical Physics*, **66**(7), 3039–3044.

286. **Unke, O. T.**, **S. Chmiela**, **H. E. Sauceda**, *et al.* (2020). Machine learning force fields. *Chemical Reviews*, **120**(21), 10718–10750.

287. **Urnov, F. D.**, **E. J. Rebar**, **M. C. Holmes**, **H. S. Zhang**, and **P. D. Gregory** (2010). Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics*, **11**(9), 636–646.

288. **Uversky, V. N.**, **J. R. Gillespie**, and **A. L. Fink** (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**(3), 415–427. ISSN 0887-3585 (Print).

289. **Vakulskas, C. A.**, **D. P. Dever**, **G. R. Rettig**, **R. Turk**, **A. M. Jacobi**, **M. A. Collingwood**, **N. M. Bode**, **M. S. McNeill**, **S. Yan**, **J. Camarena**, **C. M. Lee**, **S. H. Park**, **V. Wiebking**, **R. O. Bak**, **N. Gomez-Ospina**, **M. Pavel-Dinu**, **W. Sun**, **G. Bao**, **M. H. Porteus**, and **M. A. Behlke** (2018*a*). A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nature Medicine*, **24**(8), 1216–1224. ISSN 1546-170X. URL `https://doi.org/10.1038/s41591-018-0137-0`.

290. **Vakulskas, C. A.**, **D. P. Dever**, **G. R. Rettig**, **R. Turk**, **A. M. Jacobi**, **M. A. Collingwood**, **N. M. Bode**, **M. S. McNeill**, **S. Yan**, **J. Camarena**, *et al.* (2018*b*). A high-fidelity cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nature medicine*, **24**(8), 1216–1224.

291. **Valdés-Tresanco, M. S.**, **M. E. Valdés-Tresanco**, **P. A. Valiente**, and **E. Moreno** (2021). gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *Journal of Chemical Theory and Computation*, **17**(10), 6281–6291. ISSN 1549-9618. URL `https://doi.org/10.1021/acs.jctc.1c00645`.

292. **van der Oost, J.**, **E. R. Westra**, **R. N. Jackson**, and **B. Wiedenheft** (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature reviews. Microbiology*, **12**(7), 479–492. ISSN 1740-1534 (Electronic).

293. **van Haasteren, J.**, **J. Li**, **O. J. Scheideler**, **N. Murthy**, and **D. V. Schaffer** (2020). The delivery challenge: fulfilling the promise of therapeutic genome editing. *Nature biotechnology*, **38**(7), 845–855.

294. **van Kampen, S. J.** and **E. van Rooij** (2019). Crispr craze to transform cardiac biology. *Trends in molecular medicine*, **25**(9), 791–802.

295. **Vanommeslaeghe, K.**, **E. Hatcher**, **C. Acharya**, **S. Kundu**, **S. Zhong**, **J. Shim**, **E. Darian**, **O. Guvench**, **P. Lopes**, **I. Vorobyov**, *et al.* (2010). Charmm general force field: A force field for drug-like molecules and its application to protein-ligand interactions. *Journal of Computational Chemistry*, **31**(4), 671–690.

296. **Verlet, L.** (1967). Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, **159**(1), 98–103.

297. **Walton, R. T.**, **K. A. Christie**, **M. N. Whittaker**, and **B. P. Kleinstiver** (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science (New York, N.Y.)*, **368**(6488), 290–296. ISSN 1095-9203 (Electronic).

298. **Wan, H.**, **S. Chang**, **J.-p. Hu**, **Y.-x. Tian**, and **X.-h. Tian** (2015). Molecular Dynamics Simulations of Ternary Complexes: Comparisons of LEAFY Protein Binding to Different DNA Motifs. *Journal of chemical information and modeling*, **55**(4), 784–794. ISSN 1549-960X (Electronic).

299. **Wan, H.**, **J.-p. Hu**, **X.-h. Tian**, and **S. Chang** (2013). Molecular dynamics simulations of wild type and mutants of human complement receptor 2 complexed with C3d. *Phys. Chem. Chem. Phys.*, **15**(4), 1241–1251. URL `http://dx.doi.org/10.1039/C2CP41388D`.

300. **Wan, W. e. a.** (2019). Therapeutic approaches for targeting rna in neurodegenerative diseases. *Frontiers in Molecular Biosciences*, **6**, 96. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7036953/`.

301. **Wang, J.**, **R. M. Wolf**, **J. W. Caldwell**, **P. A. Kollman**, and **D. A. Case** (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, **25**(9), 1157–1174.

302. **Wang, X.**, **Y. Wang**, **X. Wu**, **J. Wang**, **Y. Wang**, **Z. Qiu**, **T. Chang**, **H. Huang**, **R. Lin**, and **J. Yee** (2015). Unbiased detection of off-target cleavage by crispr-cas9 and talens using integrase-defective lentiviral vectors. *Nature Biotechnology*, **33**, 175–178. CrossRef.

303. **Wang, Y.**, **X. Cheng**, **Q. Shan**, **Y. Zhang**, **J. Liu**, **C. Gao**, and **J. Qiu** (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology*, **32**, 947–951. CrossRef, PubMed.

304. **Wang, Y.**, **J. Mallon**, **H. Wang**, **D. Singh**, **M. H. Jo**, **B. Hua**, **S. Bailey**, and **T. Ha** (2021). Real-time observation of Cas9 postcatalytic domain motions. *Proceedings of the National Academy of Sciences*, **118**(2), e2010650118. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2010650118`.

305. **Watters, K. E.**, **C. Fellmann**, **H. B. Bai**, **S. M. Ren**, and **J. A. Doudna** (2018). Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science*, **362**(6411), 236–239. ISSN 10959203.

306. **Weiner, P. K.**, **P. A. Kollman**, **D. A. Case**, **U. C. Singh**, **C. Ghio**, **G. Alagona**, **S. Profeta**, and **P. Weiner** (1986). Amber: A package of computer programs for applying molecular mechanics, normal mode analysis, and free energy calculations to simulate the structures and properties of molecules. *Journal of Computational Chemistry*, **7**(2), 230–252.

307. **Wienert, B.**, **S. Wyman**, **C. Yeh**, **B. Conklin**, and **J. Corn** (2020). Crispr off-target detection with discover-seq. *Nature Protocols*, **15**, 1775–1799. CrossRef.

308. **Wong, N.**, **W. Liu**, and **X. Wang** (2015). Wu-crispr: Characteristics of functional guide rnas for the crispr/cas9 system. *Genome Biology*, **16**, 218. CrossRef, PubMed.

309. **Wu, X.**, **A. J. Kriz**, and **P. A. Sharp** (2014). Target specificity of the CRISPR-Cas9 system. *Quantitative biology (Beijing, China)*, **2**(2), 59–70. ISSN 2095-4689 (Print).

310. **Wu, Y.**, **Y. Liu**, **X. Lv**, **J. Li**, **G. Du**, and **L. Liu** (2020). Applications of CRISPR in a Microbial Cell Factory: From Genome Reconstruction to Metabolic Network Reprogramming. *ACS Synthetic Biology*, **9**(9), 2228–2238. URL `https://doi.org/10.1021/acssynbio.0c00349`.

311. **Xiaorae4, A.**, **Z. Cheng**, **L. Kong**, **Z. Zhu**, **S. Lin**, **G. Gao**, and **B. Zhang** (2014). Casot: A genome-wide cas9/grna off-target searching tool. *Bioinformatics*, **30**, 1180–1182.

312. **Yamada, M.**, **Y. Watanabe**, **J. S. Gootenberg**, **H. Hirano**, **F. A. Ran**, **T. Nakane**, **R. Ishitani**, **F. Zhang**, **H. Nishimasu**, and **O. Nureki** (2017). Crystal Structure of the Minimal Cas9 from Campylobacter jejuni Reveals the Molecular Diversity in the CRISPR-Cas9 Systems. *Molecular cell*, **65**(6), 1109–1121.e3. ISSN 1097-4164 (Electronic).

313. **Yang, H.**, **S. Ren**, **S. Yu**, **H. Pan**, **T. Li**, **S. Ge**, **J. Zhang**, and **N. Xia** (2020). Methods favoring homology-directed repair choice in response to crispr/cas9 induced-double strand breaks. *International Journal of Molecular Sciences*, **21**(18), 6461.

314. **Yang, M.**, **S. Peng**, **R. Sun**, **J. Lin**, **N. Wang**, and **C. Chen** (2018). The conformational dynamics of cas9 governing dna cleavage are revealed by single-molecule fret. *Cell reports*, **22**(2), 372–382.

315. **Yin, H.**, **C. Song**, **S. Suresh**, **S. Kwan**, **Q. Wu**, **S. Walsh**, **J. Ding**, **R. Bogorad**, **L. Zhu**, and **S. Wolfe** (2018). Partial dna-guided cas9 enables genome editing with reduced off-target activity. *Nature Chemical Biology*, **14**, 311–316.

316. **Yip, B. H.** (2020). Recent advances in crispr/cas9 delivery strategies. *Biomolecules*, **10**(6), 839.

317. **Yoon, H.**, **L. N. Zhao**, and **A. Warshel** (2019). Exploring the Catalytic Mechanism of Cas9 Using Information Inferred from Endonuclease VII. *ACS Catalysis*, **9**(2), 1329–1336. URL https://doi.org/10.1021/acscatal.8b04324.

318. **Yun, Y.** and **Y. Ha** (2020). Crispr/cas9-mediated gene correction to understand als. *International journal of molecular sciences*, **21**(11), 3801.

319. **Zeng, X.**, **Y.-W. Chen**, and **C. Tao** (2009). Feature Selection Using Recursive Feature Elimination for Handwritten Digit Recognition, 1205–1208.

320. **Zgarbov, M.**, **M. Otyepka**, **J. Sponer**, **A. Mladek**, **P. Banas**, **T. E. Cheatham III**, and **P. Jurecka** (2011). 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*, **7**(9), 2886–2902.

321. **Zgarbová, M.** and **et al.** (2015). Refinement of the sugar–phosphate backbone torsion beta for amber force fields improves the description of z-and b-dna. *Journal of chemical theory and computation*, **11**(12), 5723–5736.

322. **Zhang, D.**, **H. Zhang**, **T. Li**, **K. Chen**, **J.-L. Qiu**, and **C. Gao** (2017). Perfectly matched 20-nucleotide guide RNA sequences enable robust genome editing using high-fidelity SpCas9 nucleases. *Genome Biology*, **18**(1), 191. ISSN 1474-760X. URL https://doi.org/10.1186/s13059-017-1325-9.

323. **Zhang, G.**, **Z. Dai**, and **X. Dai** (2020*a*). C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. *Computational and Structural Biotechnology Journal*, **18**, 344–354. ISSN 2001-0370. URL https://www.sciencedirect.com/science/article/pii/S2001037019303186.

324. **Zhang, J.**, **Y. Zhang**, **S. Khanal**, **D. Cao**, **J. Zhao**, **X. Dang**, **L. N. T. Nguyen**, **M. Schank**, **X. Y. Wu**, **Y. Jiang**, *et al.* (2023). Synthetic grna/cas9 ribonucleoprotein targeting hbv dna inhibits viral replication. *Journal of Medical Virology*, **95**(7), e28952.

325. **Zhang, L.**, **H. T. Rube**, **C. A. Vakulskas**, **M. A. Behlke**, **H. J. Bussemaker**, and **M. A. Pufall** (2020*b*). Systematic in vitro profiling of off-target affinity, cleavage and efficiency for crispr enzymes. *Nucleic acids research*, **48**(9), 5037–5053.

326. **Zhang, X.-H.**, **L. Y. Tee**, **X.-G. Wang**, **Q.-S. Huang**, and **S.-H. Yang** (2015*a*). Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular therapy. Nucleic acids*, **4**(11), e264. ISSN 2162-2531 (Print).

327. **Zhang, X.-H.**, **L. Y. Tee**, **X.-G. Wang**, **Q.-S. Huang**, and **S.-H. Yang** (2015*b*). Off-target effects in crispr/cas9-mediated genome engineering. *Molecular Therapy-Nucleic Acids*, **4**.

328. **Zhang, Y.**, **H. Zhang**, **X. Xu**, **Y. Wang**, **W. Chen**, **Y. Wang**, **Z. Wu**, **N. Tang**, **Y. Wang**, **S. Zhao**, *et al.* (2020*c*). Catalytic-state structure and engineering of streptococcus thermophilus cas9. *Nature Catalysis*, **3**(10), 813–823.

329. **Zhang, Y.** *et al.* (2022). Rna modulation in cancer therapy. *Cell Death and Disease*, **13**, 540. URL https://www.nature.com/articles/s41419-022-05075-2.

330. **Zhao, Y. e. a.** (2021). Small molecule targeting of rna with high selectivity. *Advanced Science*, **8**(7), 20685–20694. URL https://onlinelibrary.wiley.com/doi/full/10.1002/advs.202004685.

331. **Zheng, L.**, **J. Shi**, and **Y. Mu** (2018). Dynamics changes of CRISPR-Cas9 systems induced by high fidelity mutations. *Phys. Chem. Chem. Phys.*, **20**(43), 27439–27448. URL http://dx.doi.org/10.1039/C8CP04226H.

332. **Zhu, H.**, **C. Li**, and **C. Gao** (2020). Applications of crispr–cas in agriculture and plant biotechnology. *Nature Reviews Molecular Cell Biology*, **21**(11), 661–677.

333. **Zhu, X.**, **R. Clarke**, **A. K. Puppala**, **S. Chittori**, **A. Merk**, **B. J. Merrill**, **M. Simonović**, and **S. Subramaniam** (2019). Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nature Structural & Molecular Biology*, **26**(8), 679–685. ISSN 1545-9985. URL https://doi.org/10.1038/s41594-019-0258-2.

334. **Zok, T.**, **M. Antczak**, **M. Zurkowski**, **M. Popenda**, **J. Blazewicz**, **R. W. Adamiak**, and **M. Szachniuk** (2018). Rnapdbee 2.0: multifunctional tool for rna structure annotation. *Nucleic acids research*, **46**(W1), W30–W35.

335. **Zuo, Z.** and **J. Liu** (2016). Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations. *Scientific Reports*, **6**(1), 37584. ISSN 2045-2322. URL https://doi.org/10.1038/srep37584.

336. **Zuo, Z.** and **J. Liu** (2017). Structure and Dynamics of Cas9 HNH Domain Catalytic State. *Scientific Reports*, **7**(1), 17271. ISSN 2045-2322. URL https://doi.org/10.1038/s41598-017-17578-6.

337. **Zuo, Z.** and **J. Liu** (2020). Allosteric regulation of CRISPR-Cas9 for DNA-targeting and cleavage. *Current Opinion in Structural Biology*, **62**, 166–174. ISSN 0959-440X. URL https://www.sciencedirect.com/science/article/pii/S0959440X20300130.