Structure-based thermodynamics of PAM Recognition by CRISPR/Cas9: Insight from computer simulations

A Thesis

Submitted in partial fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

Shreya Bhattacharya

Department of Biosciences and Bioengineering

Indian Institute of Technology

Guwahati 781039, India

Dedicated to my family and friends

Indian Institute of Technology Guwahati
Department of Biosciences and Bioengineering

DECLARATION

I hereby declare that the content of this thesis entitled "**Structure-based thermodynamics of PAM recognition by CRISPR/Cas9: Insight from computer simulations**" is an original work carried out by me in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, for the award of the degree of Doctor of Philosophy, under the supervision of Dr. Priyadarshi Satpati.

I further declare that this thesis has not been submitted previously, in part or in full, for the award of any degree, diploma, or any other academic qualification at this or any other institution.

As per standard scientific practice, due acknowledgement has been made wherever the work described is based on the findings of other researchers.

Shreya Bhattacharya

Date: 28/01/2026

Place: Guwahati

Shreya Bhattacharya

Roll no. 216106013

Indian Institute of Technology Guwahati
Department of Biosciences and Bioengineering

CERTIFICATE

This is to certify that the thesis entitled "**Structure-based thermodynamics of PAM recognition by CRISPR/Cas9: Insight from computer simulations**", submitted by **Ms. Shreya Bhattacharya** (Roll No. 216106013) for the award of the degree of Doctor of Philosophy, is a bona fide record of the research work carried out by her under my supervision in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India.

The work embodied in this thesis has not been submitted elsewhere for the award of any degree or diploma.

Date: 28/01/2026

Place: Guwahati

Dr. Priyadarshi Satpati

(Thesis Supervisor)

Acknowledgement

I extend my sincere gratitude to **Dr. Priyadarshi Satpati**, my PhD supervisor, whose invaluable guidance, insight, and encouragement have shaped every stage of this thesis. His patience, encouragement, constructive criticism, and constant support have been instrumental throughout my research journey. His feedback, sometimes gentle, sometimes very direct and endless intellectual discussions have always pushed me toward better thinking and better science. I am genuinely grateful for the freedom and trust he gave me throughout this journey and consider myself fortunate to have worked under his mentorship.

I sincerely thank my thesis examiners, **Dr. Alexey Aleksandrov** (Ecole Polytechnique, France) and **Prof. Md. Ehesan Ali** (Institute of Nano Science and Technology, Mohali, Punjab), for their careful evaluation, insightful comments, and generous appreciation, which greatly helped in refining the presentation of this work. I am also thankful to **my doctoral committee members**, Prof. Manish Kumar (Chairperson), Prof. Shankar Prasad Kanaujia and Dr. Sunanda Chatterjee for their valuable feedback, time, and expertise, which enriched the quality of this work and my understanding of the field. I further thank the **viva voce committee members**, Dr Kapish Gupta and Dr. Amit Kumar, for their constructive questions, and engaging discussions during my PhD defence.

I extend my appreciation to the **Heads of the Department**, **faculty members, research staff, and administrative team of the Department of Biosciences and Bioengineering (BSBE)** for creating a supportive academic environment and for their administrative support and providing me the necessary infrastructure to fulfil my PhD thesis objectives.

In addition, I would like to acknowledge the **Bioinformatics Infrastructure Facility** (BIF), Department of BSBE, and the **PARAM ISHAN** and **PARAM KAMRUPA** supercomputing facilities at IIT Guwahati for providing essential computational resources.

I am deeply grateful to the Ministry of Education's **Prime Minister's Research Fellowship** (PMRF) and IIT Guwahati for their financial support throughout my Ph.D.

I gratefully appreciate the support of **my friends** (Arisha, Kangkana, Eena, Sawna and others) in IIT Guwahati for their encouragement, moral support, motivation, and companionship during this journey.

I also want to acknowledge the natural beauty of IIT Guwahati campus, its serene landscapes, peaceful lakes, and the diverse animals and birds, whose quiet presence offered calm and inspiration during challenging times.

I thank **my family**, especially **my parents**, for their boundless love, belief in me, constant emotional support, and for standing by me through every challenge. Nothing in this thesis would exist without you.

Finally, I express my gratitude to everyone who contributed, directly or indirectly, to the successful completion of this thesis.

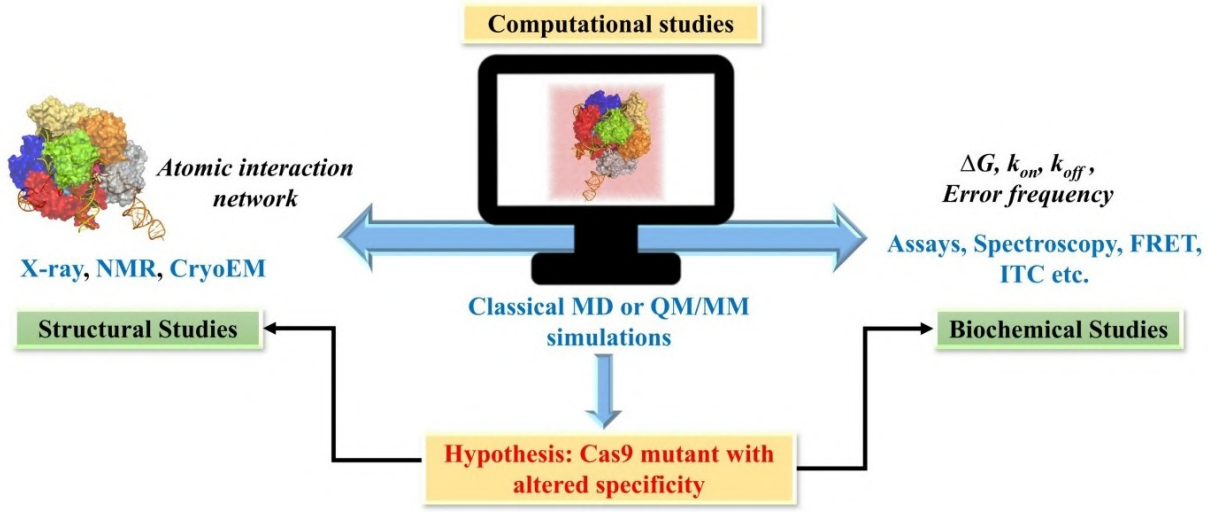
Thank you!

Shreya Bhattacharya

Synopsis

The CRISPR/Cas9 system derived from *Streptococcus pyogenes* (*Sp*Cas9) has revolutionized molecular biology by allowing precise and programmable editing of DNA sequences in living cells. *Sp*Cas9 is a multi-domain RNA-guided DNA endonuclease that uses a single guide RNA (sgRNA) to bind and cut DNA at locations adjacent to a protospacer adjacent motif (PAM), which consists of the three-nucleotide canonical sequence 5'-NGG-3' (where N can be any nucleotide). The stringent PAM requirement (5'-NGG) limits the range of genomic sites accessible for editing. Therefore, understanding the molecular and energetic basis of PAM recognition is of paramount importance for the rational engineering of Cas9 variants with broadened or altered PAM specificities. Although mutations in *Sp*Cas9 have been shown to enhance PAM recognition, the energetics of PAM recognition related to those mutations and their connection to atomic structure remain unknown. This thesis employs molecular simulations using precatalytic *Sp*Cas9:sgRNA:dsDNA as a template to clarify the structure-based free energy landscape related to PAM selectivity in *Sp*Cas9 and its engineered variants. Using alchemical free energy calculations, the research examines how amino acid mutations in *Sp*Cas9 affect DNA binding (discussed in chapters 2 and 3). Additionally, chapter 4 explores how the PAM binding affinity of *Sp*Cas9 changes in response to mutations in the canonical 5'-NGG sequence. Results indicated that the PAM recognition by *Sp*Cas9 is influenced by the local hydrophobicity and flexibility of its binding cleft. The flexibility of the protein residues facilitates new interactions, while the hydrophobicity enhances the interactions in non-canonical PAM sequences, thereby broadening PAM readability. The study establishes a direct connection between the estimated energetics and the molecular structures and provides an explanation for the experimentally observed cleavage activity of *Sp*Cas9. This work establishes a clear framework for understanding PAM recognition in *Sp*Cas9, laying out the groundwork for designing new CRISPR-based genome editing tools. This thesis is divided into five chapters.

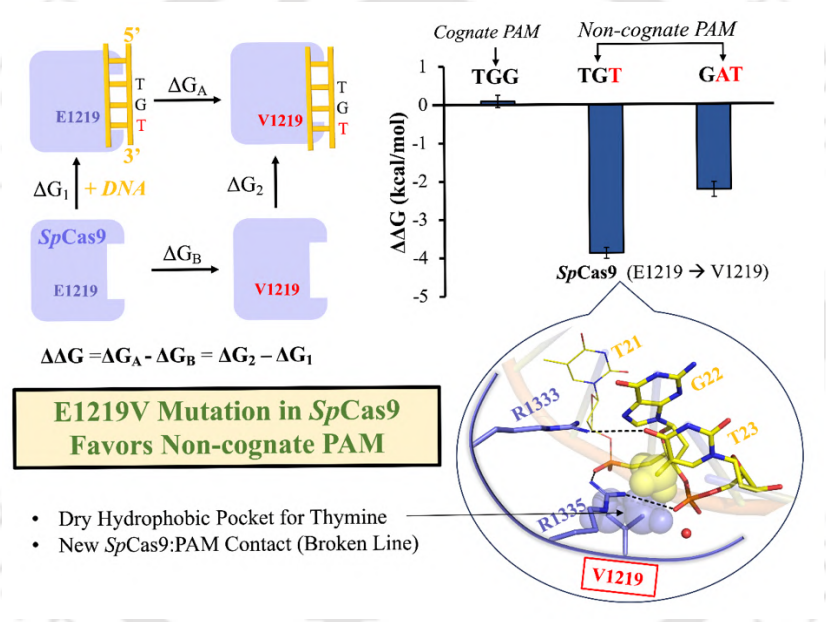
Chapter 1 presents the general introduction to the CRISPR/Cas9 system with a detailed overview of its molecular mechanism, emphasising the critical role of the protospacer adjacent motif (PAM) in target DNA recognition. The chapter presents a review of relevant structural and biochemical studies and clearly outlines the current state of knowledge. Importantly, it identifies unresolved questions concerning the energetic origins of PAM specificity and the mechanisms by which engineered Cas9 variants achieve relaxed PAM requirements. The chapter also reviews existing literature to highlight how molecular dynamics simulations have bridged the knowledge gap between structural and biochemical studies to a great extent.



The CRISPR/Cas9 system has transformed modern genome engineering, yet the molecular basis of its strict PAM requirement remains only partially understood. For *Streptococcus pyogenes* Cas9 (*Sp*Cas9), recognition of the 5′-NGG-3′ PAM is a checkpoint that determines whether DNA binding and subsequent cleavage can occur. Although structural and biochemical studies have identified key residues involved in PAM interaction, they do not fully explain the energetic origins of specificity or how engineered variants achieve broader PAM readability. This chapter establishes the biological context necessary to motivate a quantitative investigation of PAM recognition and defines the knowledge gap that drives the present work. This foundation leads to two central research questions: (i) how do protein mutations influence *Sp*Cas9:PAM binding affinity? and (ii) how does *Sp*Cas9 discriminate between canonical and non-canonical PAM

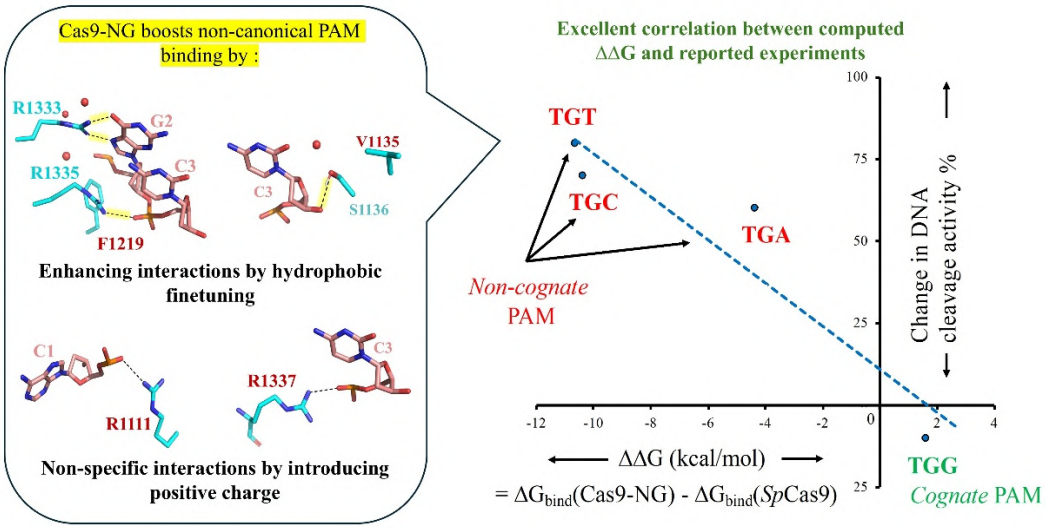
sequences? These questions formulate the research objectives that guide the investigation into an important aspect of *Sp*Cas9 specificity. The chapter concludes with a description of the computational methodology employed in this thesis, in particular classical molecular dynamics simulations and alchemical free-energy calculations.

Chapter 2 estimated the change in the PAM binding affinity in response to a single E1219V mutation in *Sp*Cas9. The E1219V mutation in *Sp*Cas9 fine-tunes the water accessibility in the PAM binding pocket and promotes new interactions in the *Sp*Cas9: non-canonical T-rich PAM, thus weakening the PAM stringency.



The nucleotide-specific interaction of two arginine residues (i.e., R1333 and R1335 of *Sp*Cas9) ensured stringent 5′-NGG-3′ PAM recognition. R1335A substitution (*Sp*Cas9^{R1335A}) completely disrupts the direct interaction between *Sp*Cas9 and PAM sequences (canonical or noncanonical), accounting for the loss of editing activity. Interestingly, the double mutant (*Sp*Cas9^{R1335A,E1219V}) boosts DNA binding affinity by favouring protein:PAM electrostatic contact in a desolvated pocket. The underlying thermodynamics explains the varied DNA cleavage activity of *Sp*Cas9 variants. A direct link between the energetics, structures, and activity is highlighted, which can aid in the rational design of improved *Sp*Cas9-based genome editing tools.

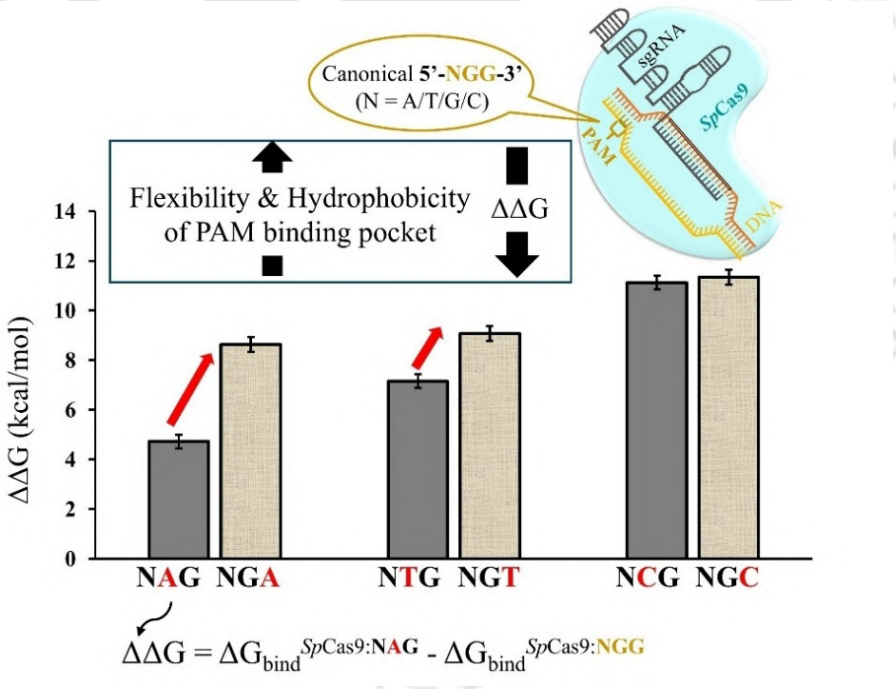
Chapter 3 explored the energetic basis for the enhanced PAM readability in engineered Cas9-NG (*Sp*Cas9 with seven mutations: R1335V, E1219F, D1135V, L1111R, T1337R, G1218R, and A1322R). The changes in PAM binding affinity (TGG, TGA, TGT, or TGC) for each of the seven *Sp*Cas9 mutations, were calculated based on more than 53 μs of MD. The underlying thermodynamics (ΔΔG) accounts for the experimentally observed differences in DNA cleavage activity between *Sp*Cas9 and Cas9-NG across various DNA substrates.



The interaction energies between *Sp*Cas9 and DNA are significantly influenced by the type and location of the amino acid mutations. Notably, the R1335V mutation disfavors DNA binding by disrupting critical interactions with the PAM. However, the destabilizing effect of the R1335V mutation is mitigated by four advantageous mutations (E1219F, D1135V, L1111R, and T1337R), which primarily introduce non-base-specific interactions and enhance PAM readability. The hydrophobic substitutions (E1219F and D1135V) are particularly impactful, as they exclude solvent from the PAM binding pocket, strengthening electrostatic interactions in the low dielectric medium and increasing the stability of the noncognate PAM complexes by ~2-5 kcal/mol. Additionally, L1111R and T1337R facilitate DNA binding by forming direct electrostatic contacts. In contrast, the charge mutations G1218R and A1322R do not effectively promote interactions with the negatively charged DNA, clearly demonstrating that the spatial location of mutations is crucial in shaping this interaction energetics. We demonstrated that stabilization of the Cas9-NG:

noncognate PAM complexes enables broader PAM recognition. This is primarily achieved through two mechanisms: (1) the establishment of new non-base-specific interactions between the protein and nucleotides and (2) the enhancement of electrostatic interactions within a relatively dry and hydrophobic pocket. The findings revealed that mutation-induced desolvation can improve the recognition of noncognate PAMs, paving the way for the rational and innovative design of *Sp*Cas9 mutants.

Chapter 4 evaluated the strength of PAM recognition by wild-type *Sp*Cas9. These calculations determine the change in *Sp*Cas9 binding affinity (ΔΔG) due to base mutations at all three positions of 5'-NGG within the pre-catalytic complex.



*Sp*Cas9 does not discriminate at the first position of the consensus NGG sequence, but it penalizes mutations in the second and third positions. We demonstrate that *Sp*Cas9 imposes a 3 kcal/mol higher penalty for guanine mutation in the third PAM position compared to the second. This difference is due to the greater conformational rigidity of R1335 in relation to R1333. The penalty of the third guanine mutation increases significantly because the rigid R1335 cannot readjust and

form new protein-DNA interactions to compensate for the missing interactions in the noncanonical sequence. A cytosine-to-guanine substitution in either the second or third position of canonical PAM disrupts direct protein-PAM interactions and leads to solvent exposure. This happens due to strong electrostatic repulsion between the arginine dyad's guanidine groups and the amine group of cytosine. Interestingly, single cytosine substitutions can be penalized (by more than 10 kcal/mol) as strongly as multiple guanine substitutions in the NGG sequence, reflecting the sensitivity of the arginine dyad to electrostatic repulsion and solvent effects. The ability of *Sp*Cas9 to differentiate between non-canonical and canonical PAMs (ΔΔG) is directly related to the number of direct interactions between *Sp*Cas9 and the PAM sequence, as well as the degree of solvent exposure. A decrease in direct interactions, combined with increased solvent exposure, leads to an enhancement of ΔΔG. The calculated ΔΔG adequately explains the observed differences in DNA cleavage activity of *Sp*Cas9 across various DNA substrates with different PAM sequences. This study connects thermodynamics, structures, and activity to elucidate PAM selectivity in *Sp*Cas9 and may also be applicable to other CRISPR/Cas systems, offering valuable insights for the rational design of Cas9 variants with modified PAM specificities.

Chapter 5 presents the overall conclusions of this thesis and outlines the future directions for extending the current work to other Cas family proteins. By integrating molecular dynamics simulations with alchemical free-energy calculations, the work establishes how electrostatics, solvation effects, and residue flexibility shape the recognition landscape of *Sp*Cas9 and its engineered variants. We believe the methodology employed and the hypothesis formulated in this study will contribute meaningfully to the rational design of enhanced *Sp*Cas9 variants with altered PAM specificity.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Form
ATP	Adenosine triphosphate
BAR	Bennett Acceptance Ratio
bp	Base pairs
Cas9	CRISPR-associated protein 9
CHARMM36	Chemistry at HARvard Macromolecular Mechanics (version 36)
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
CryoEM	Cryo-Electron Microscopy
DNA	Deoxyribonucleic Acid
dsDNA	Double-stranded DNA
DSB	Double-stranded break
FEP	Free Energy Perturbation
GFP	Green Fluorescent Protein
GROMACS	GROningen MAchine for Chemical Simulations
HDR	Homology Directed Repair
HIV	Human immunodeficiency virus
HNH	Nuclease domain in Cas9
MD	Molecular Dynamics
MM/GBSA	Molecular mechanics/generalized Born surface area
MM/PBSA	Molecular Mechanics/Poisson–Boltzmann Surface Area
NAMD	NAnoscale Molecular Dynamics
NHEJ	Non-homologous end-joining
NGG	Canonical PAM sequence
NRN	PAM recognition pattern (Purine-any-Purine)
NYN	PAM recognition pattern (Pyrimidine-any-Pyrimidine)
nt	Nucleotides

ntDNA	Nontarget DNA
NUC	Nuclease domain
PAM	Protospacer Adjacent Motif
PCA	Principal Component Analysis
PDB	Protein Data Bank
PBC	Periodic Boundary Conditions
PDB	Protein Data Bank
PI	PAM Interacting domain
PID	PAM Interacting Domain
PSF	Protein Structure File
PyMOL	Python-based Molecular Viewer
QM/MM	Quantum mechanics/molecular mechanics
REC	Recognition domain
Rg	Radius of gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNA	Ribonucleic Acid
RNP	Ribonucleoprotein
RuvC	Nuclease domain in Cas9
*Sa*Cas9	*Staphylococcus aureus* Cas9
SASA	Solvent accessible surface area
sgRNA	Single guide RNA
*Sp*Cas9	*Streptococcus pyogenes* Cas9
TALENs	Transcription activator-like effector nucleases
tDNA	Target DNA
TI	Thermodynamic Integration
TIP3P	Three-site transferable intermolecular potential water model
tracrRNA	Tracer RNA
VMD	Visual Molecular Dynamics
ZFNs	Zinc-finger nucleases

Symbols

Mg^{2+}	Magnesium ions	kcal/mol	Kilocalorie per mole	V	Volume
ΔG	Gibbs free energy	kDa	Killo Dalton	F	Force
ΔΔG	Relative binding free energy	Å	Angstrom	λ	Coupling coordinate
t	time	nm	Nanometer		
fs	Femtoseconds	T	Temperature		
ns	Nanoseconds	°C	Degree Celsius		
ps	Picoseconds	K	Kelvin		
μs	Microsecond	n	Number of atoms		
		v	Velocity		

Chapter 1

Introduction, Motivation, Objectives and Methodology

Part of this chapter is published in ACS Omega, 2022, 8, 2, 1817–1837

1.1 Introduction

CRISPR/Cas9 ("Clustered Regularly Interspaced Short Palindromic Repeats" and "CRISPR-associated protein Cas9") technology is of significant interest for genome editing (Pickar-Oliver and Gersbach, 2019; Ray et al., 2019). CRISPR/Cas9 is a part of the adaptive immunity of bacteria and archaea, which eliminates foreign genetic material (viz., invading bacteriophages) (Pickar-Oliver & Gersbach, 2019) and is repurposed as a groundbreaking technique that allows scientists to edit regions of the genome by deleting, inserting, or modifying DNA sequences. CRISPR/Cas9 is a simple two-component system that consists of RNA component: (crRNA and tracrRNA or sgRNA) and a single protein component called Cas9. Engineering the crRNA only is sufficient to target particular DNA sequences using the CRISPR/Cas9 tool (Barrangou & Doudna, 2016; Deltcheva et al., 2011). Thus, contrary to other protein-guided genome editing tools like ZFNs and TALENs, which require intensive protein engineering, CRISPR/Cas9 is a simple, affordable, and, more importantly, efficient tool for genome editing (Tiruneh et al., 2021). Application of CRISPR/Cas9 technology can introduce site-specific mutations, knockout disease-causing genes, specific gene knock-ins, change gene expression levels, and many more (Tiruneh et al., 2021). These can be accomplished by using a single guide RNA (sgRNA: crRNA fused to the tracrRNA by a linker) that has exact complementarity to the gene of interest and a Cas9 endonuclease that cleaves the specific sequence of interest (Barrangou & Doudna, 2016; Yin et al., 2016).

One of the recent breakthroughs in CRISPR/Cas9 technology was the development of the CRISPR-based drug CTX001 to cure sickle cell disease (SCD) and β-Thalassemia (Frangoul et

al., 2024; Pattan et al., 2021). In 2023, the first CRISPR-based gene-editing therapy, Casgevy (exagamglogene autotemcel), a CRISPR-based therapy for SCD and β-thalassemia, which corrects genetic defects by editing hematopoietic stem cells (Frangoul et al., 2024), received regulatory approvals in the UK and the US for the treatment of SCD and transfusion-dependent β-thalassemia, marking a major milestone in the field (U.S. FDA, 2023). Recent clinical trials have further demonstrated CRISPR/Cas9's efficacy in treating other genetic disorders, such as transthyretin amyloidosis, marking a new era in precision medicine (Gillmore et al., 2021). Thus, CRISPR-based technology appears to be a promising approach for correcting genetic defects in the coming decades (Braga et al., 2022).

Advancements in the structural study and the development of cutting-edge computer technology with ever-increasing computational power have opened up the possibility of performing structure-based computational analysis (Karplus & McCammon, 2002; Sharma et al., 2022). Molecular Dynamics (MD) simulations have been a popular method for understanding the dynamics of biomolecules at a molecular level by extensive exploration of phase space (Karplus & McCammon, 2002). Such studies have boosted knowledge of the molecular mechanism of CRISPR/Cas9 in terms of structures, thermodynamics, and kinetics and have served as an excellent complement to experimental studies. MD simulations of the CRISPR/Cas9 system have revealed several key aspects, including nucleotide binding (Nierzwicki et al., 2021; Palermo et al., 2016; Palermo, Miao, et al., 2017), catalytic mechanism (Ahsan et al., 2023; Casalino et al., 2020; Palermo, 2019; Zuo & Liu, 2016, 2017), and off-target effects (Arantes et al., 2023; Bhardwaj et al., 2024; Mitchell et al., 2020; Ricci et al., 2019). These studies complement experimental approaches and rapidly transform the understanding of CRISPR/Cas9 activity at a fundamental level, enabling the design of CRISPR systems with improved specificity and efficiency.

1.2. Discovery and origin of CRISPR/Cas systems

The timeline of important milestones related to CRISPR is given in **Figure 1.1**. The first description of CRISPR loci emerged in 1987, when Nakata et al. identified an interesting locus downstream to the IAP gene (encoding Alkaline phosphatase Isozyme) of *E. coli* that contain roughly palindromic repeated sequences (direct repeat of 29 bp) inter-spaced by variable

sequences (Nakata et al., 1987). In the subsequent years, with the rapid advancement in genome sequencing, researchers started identifying similar sequences in bacteria (Lander, 2016; Warren et al., 2002; Chen et al., 2005; Altermann et al., 2005), and archaea (Lander, 2016). The name CRISPR as "Clustered Regularly Interspaced Short Palindromic Repeat" was first coined by Jansen et al. in 2002, who showed that apart from the CRISPR array containing variable spacers and short palindromic repeats, there is a set of "Cas" or "CRISPR associated genes" constantly associated with CRISPR loci (Jansen et al., 2002). In 2005, three independent computational studies (**Figure 1.1**) reported that the spacer sequences between the repeats identified in the bacteria actually match with the foreign DNA and specifically bacteriophage DNA (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). Thus, the spacers were hypothesized to act as a memory of viral infection and provide cellular immunity against phage infections (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). Barrangou et al. (Barrangou et al., 2007) provided the first experimental proof in 2007 that the CRISPR system was involved in bacterial immunity to protect itself against foreign DNA and bacteriophages and confirmed the computational predictions (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). They infected the *Streptococcal* strain with two different bacteriophages and noticed that the bacteriophage sequences were integrated (triggering CRISPR system) in the bacterial genome and developed resistance to that bacteriophage (Barrangou et al., 2007). Brouns et al. demonstrated the process of mature crRNA formation from CRISPR loci in *E. coli* (Brouns et al., 2008). The ability of CRISPR/Cas9 tool to selectively target any DNA of interest was demonstrated by Jinek et al. in 2012. They used CRISPR/Cas9 to cleave Green Fluorescent Protein (GFP) at five specific genomic sites (Jinek et al., 2012). Additionally, they designed a single chimeric RNA (linking crRNA and tracrRNA as sgRNA) that would allow Cas9 to for cleave the target DNA using only one RNA component (Jinek et al., 2012). This discovery of Cas9 as a programmable genome editing tool (Jinek et al., 2012) made a breakthrough and led Emmanuelle Charpentier and Jennifer Doudna to win the Nobel prize in 2020. In 2013, Cong et al. successfully used CRISPR/Cas9 tool for editing the genome of human and mouse cell cultures (Cong et al., 2013). The first crystal structure of apo Cas9 and Cas9:sgRNA:target DNA complex was revealed in 2014 by Jinek et al (Jinek et al., 2014) and Nishimasu et al.,(Nishimasu et al., 2014) respectively. The application of the first Molecular Dynamics (MD) simulation in CRISPR system was first

reported in 2015 by Estarellas et al., where, they studied the dynamics of CRISPR/Csy4 complex (Estarellas et al., 2015). Csy4 or Cas6f is a Cas-protein involved in the maturation of pre-crRNA to crRNA (Haurwitz et al., 2010). The first therapeutic application of CRISPR/Cas9 in human zygotes was carried out by Tang et al. in 2017, where they modified mutations in the β-globin gene (Tang et al., 2017). One of the most remarkable and controversial CRISPR/Cas9 application was reported by He Jiankui in November 2018, when twin girls "Lulu and Nana" had their CCR5 gene of their embryos genetically modified to make them HIV resistant (Gouw, 2019). The girls were born to an HIV-positive father and an HIV-negative mother and were claimed to be resistant to HIV via editing of their chemokine receptor (CCR5) gene (Greely, 2019). CCR5 encodes a protein that HIV uses to invade host cells; therefore, specific mutations were introduced into the CCR5 gene that would confer innate HIV resistance to the bacteria (Greely, 2019). Although this germline gene editing in humans raised ethical concerns (Krimsky, 2019), the immense potential of CRISPR/Cas9 in curing genetic disease and thus improving human life in the future is undeniable. Clinical trials for the first in vivo application of Cas9 began in 2019 (Gillmore et al., 2021). In 2023, "Casgevy" (exagamglogene autotemcel) was the first FDA approved CRISPR-based therapy for SCD and β-thalassemia, (Frangoul et al., 2024), which edits hematopoietic stem cells correcting genetic defects. This success has paved the way for ongoing research into CRISPR-based treatments for other monogenic disorders, such as cystic fibrosis and muscular dystrophy, where precise gene correction could address the root causes of disease (Mani, 2021). Presently, CRISPR/Cas9 is used in vast applications in genome editing ranging from therapeutic applications (Adane & Alamnie, 2024; Ahmed et al., 2021; Bhushan et al., 2024; Sadeqi Nezhad et al., 2021; Seok et al., 2021; Sorolla et al., 2022; W. Wang et al., 2021; Wellhausen et al., 2021), to plant (Adane & Alamnie, 2024; Deb et al., 2022; Fizikova et al., 2021; Nguyen et al., 2022) and fungal (C. Jiang et al., 2021; Liao et al., 2021; H. Zhu et al., 2024) biotechnology.

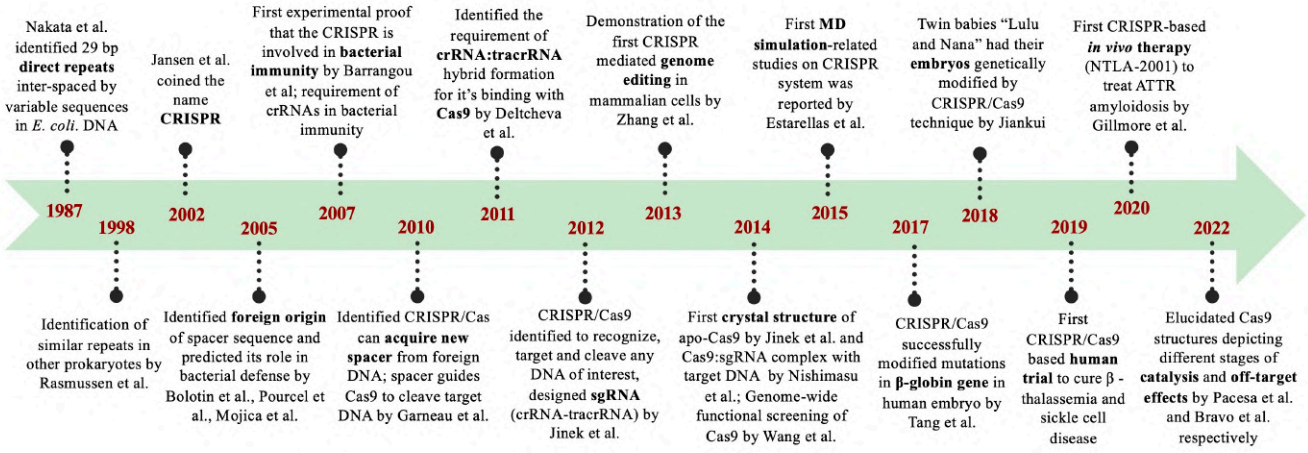


Figure 1.1. The timeline of pioneering leads related to CRISPR-Cas discovery.

1.3. Simple schematic overview of the mechanism and classification of CRISPR adaptive immunity

A simple overview of the mechanism of CRISPR/Cas9 sequence-specific adaptive immunity in most of the bacteria and archaea is given in **Figure 1.2**. CRISPR/Cas based adaptive immune system is found exclusively in prokaryotes (around 36% of bacteria and 75% of archaea) (Couvin et al., 2018; Pourcel et al., 2020). It is composed of two components: a nucleic acid component named CRISPR and protein component called Cas proteins. CRISPR locus comprises of a CRISPR array that contains short 30-40 bp direct repeats interspaced by short variable DNA sequences of viral origin (called 'spacers': S1, S2 of **Figure 1.2**). This CRISPR array is flanked by a set of CRISPR-associated (Cas) genes (Rath et al., 2015). CRISPR/Cas immunity is achieved through three stages: adaptation, expression, and interference (Amitai & Sorek, 2016; Rath et al., 2015) (**Figure 1.2**).

In the adaptation/acquisition stage, the infecting viral DNA fragment (known as 'protospacers': S0 of **Figure 1.2**) is integrated into the CRISPR locus of bacterial DNA as a new spacer sequence between a series of short repeats (Amitai & Sorek, 2016; Brouns et al., 2008; McGinn & Marraffini, 2018; Rath et al., 2015). Cas1 and Cas2 proteins are required for DNA acquisition (Amitai & Sorek, 2016; Brouns et al., 2008; McGinn & Marraffini, 2018; Rath et al., 2015).

Protospacer acquisition in CRISPR/Cas systems requires recognition of a Protospacer Adjacent Motif (PAM: **Figure 1.2**) in the viral DNA, and the DNA sequences immediately upstream of PAM are cleaved and incorporated into the CRISPR array as spacers (Amitai & Sorek, 2016; Brouns et al., 2008; McGinn & Marraffini, 2018; Rath et al., 2015). The PAM sequence is usually three-nucleotide "NGG", for most popular CRISPR/Cas9 system from *Streptococcus pyogenes* (*Sp*Cas9, as shown in **Figure 1.2**). When a new infection occurs, the spacer is always added next to the regulatory leader sequence ("L" of **Figure 1.2**), a DNA sequence facilitating the integration of spacers at the correct site (Kieper et al., 2019). Thus, the location of the spacer sequence relative to the leader sequence indicates the history of infection (spacer sequence away from the leader implies older infection) (Brouns et al., 2008; Marraffini, 2015; Rath et al., 2015). Spacers are at the center of CRISPR defense as they confer immunity against phages or plasmids that contain a complementary sequence and also provide the sequence memory for defense against subsequent invasions (Brouns et al., 2008; Marraffini, 2015; Rath et al., 2015). This sequence memory of the initial encounter of the virus helps the bacteria in evading subsequent infection from the same virus by amplifying immune response upon reinfection thus providing adaptive immunity in the bacteria.

During the expression stage, the CRISPR array is transcribed as a precursor CRISPR RNA (pre-crRNA, **Figure 1.2**). Another RNA, the tracrRNA (trans-activating CRISPR RNA, **Figure 1.2**), is also produced in the system, having a sequence complementary to the repeats (Deltcheva et al., 2011; Jiang & Doudna, 2017). The pre-crRNA forms a duplex with the tracrRNA, which serves as a scaffold for the binding of Cas proteins (usually Cas9) (Zeng et al., 2018). This duplex is then recognized by the host RNaseIII, which in the presence of Cas9 will specifically cleave to generate individual units of repeat-spacer:tracrRNA duplex or mature crRNA:tracrRNA duplex (Deltcheva et al., 2011; Jiang & Doudna, 2017), carrying one spacer sequence that guides Cas proteins. CrRNA contains an RNA sequence (~20 nt) that is complementary to the target viral DNA.

In the interference stage, Cas9 endonuclease forms a complex with crRNA:tracrRNA duplex forming a ribonucleoprotein (RNP) surveillance complex, which scans the foreign DNA of cognate viruses or plasmids in the cell (Jiang and Doudna, 2017a; Rath et al., 2015). Cas9:crRNA:tracrRNA recognizes the PAM sequence and triggers the unwinding of the foreign DNA for DNA:RNA hybridization, followed by Cas9-catalyzed specific double-stranded DNA

break (Marraffini, 2015; Rath et al., 2015), thereby neutralizing the invader by making its DNA non-functional. The absence of the PAM sequence in the CRISPR array of the bacterial genome prevents Cas9:crRNA:tracrRNA induced cleavage of its own genome (Amitai & Sorek, 2016; Brouns et al., 2008; McGinn & Marraffini, 2018; Rath et al., 2015).

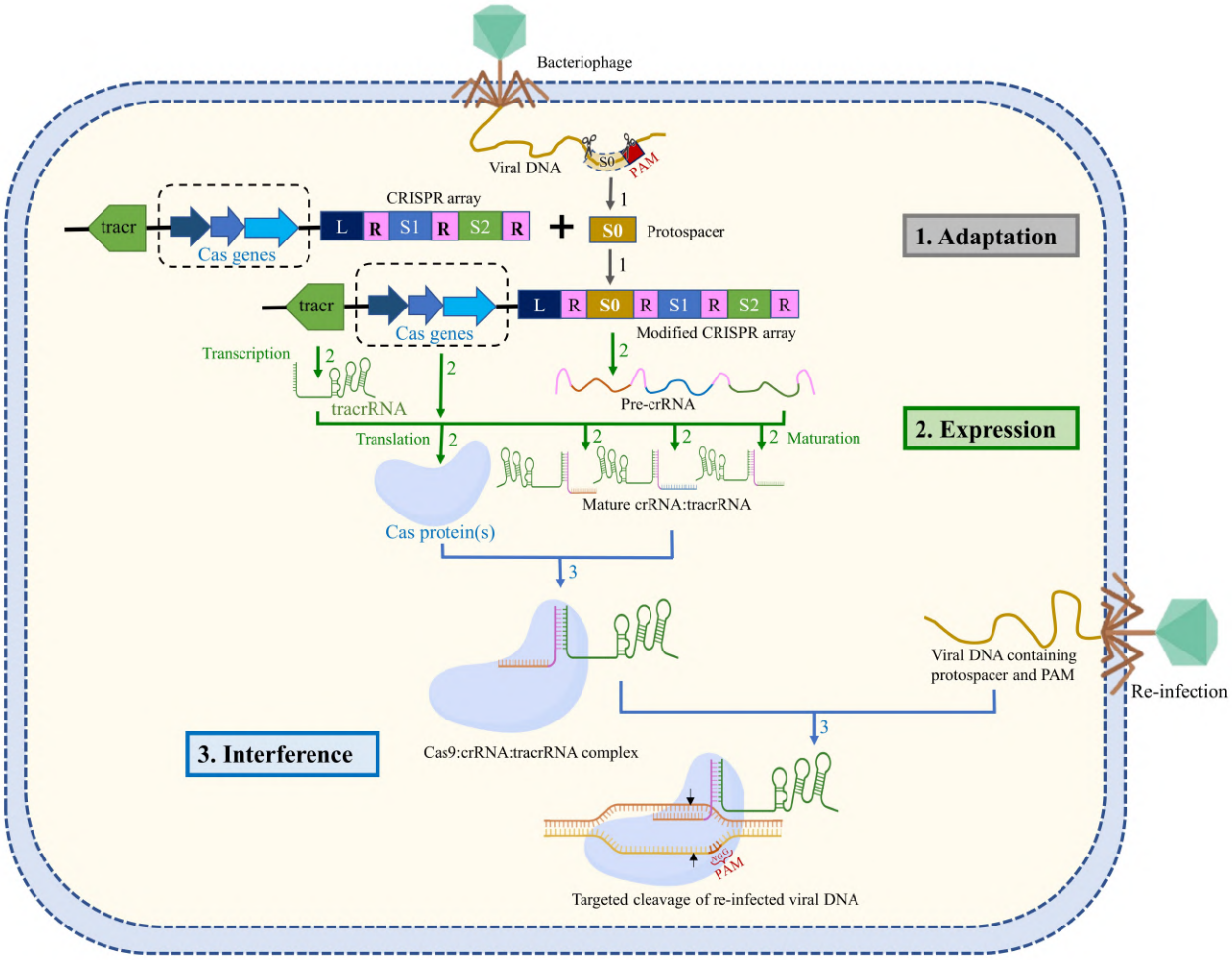


Figure 1.2. Schematic representation of the CRISPR mediated immunity acquired in bacteria through (1) **Adaptation**: Integration of foreign DNA as a spacer (S0) in CRISPR loci after leader sequence (L) flanked by direct repeats (R), (2) **Expression**: Decoding of CRISPR array into pre-crRNA, Cas protein(s) and tracrRNA followed by subsequent processing of pre-crRNA into crRNA and (3) **Interference**: Cas9 in complex with crRNA:tracrRNA neutralize the re-infection. Cas9:crRNA:tracrRNA complex recognize the foreign DNA by using PAM sequence and DNA:RNA complementarity and triggers Cas9 catalyzed DNA cleavage (cleavage site is indicated by black arrows), thus, providing immunity against viral re-infection. The target DNA (tDNA) strand is shown in orange, and the non-target strand is shown in yellow. PAM

sequence is shown in the non-target DNA (ntDNA) strand. The adaptation, expression, and interference stages are represented in grey, green and blue arrows, respectively.

1.3.1. Classification of CRISPR/Cas systems

CRISPR system includes multiple Cas proteins for its function. The rapid evolution of Cas genes and the involvement of various combinations of Cas proteins by various systems have made CRISPR/Cas classification very challenging (Makarova et al., 2011; Ray et al., 2019). CRISPR/Cas systems have been classified into two broad groups: Class I and Class II, which are further sub-classified into six types containing 33 sub-types (Koonin et al., 2017; Ray et al., 2019). The classification (Class I and Class II) is primarily based on the number of Cas proteins involved in the interference step (**Figure 1.3**). Class I systems involve multiple Cas proteins, whereas class II systems involve only one protein (viz., Cas9) for the interference step (Koonin et al., 2017; Makarova et al., 2011, 2015). CRISPR/Cas systems are also divided into different types based on the presence of unique genes (Cas3, Cas9, Cas10, Csf2, Cpf1, Cas13 for Type I to VI, respectively) (Koonin et al., 2017; Makarova et al., 2011, 2015; Taylor et al., 2021). A simple overview of the Cas protein components characteristic of each type has been depicted in **Figure 1.3**. The further classification into subtypes is more complicated, which takes into account several factors like the presence of unique genes, evolutionary conservation of effector modules, and many more (Koonin et al., 2017; Ray et al., 2019). Involvement of a single Cas protein in the interference step has made the Class II system very attractive as a genome editing tool, particularly the CRISPR/Cas9 system, which has been extensively used in numerous genome editing applications and is most widely used for genome editing. (Ahmed et al., 2021; Deb et al., 2022; Fizikova et al., 2021; C. Jiang et al., 2021; Liao et al., 2021; Nguyen et al., 2022; Sadeqi Nezhad et al., 2021; Seok et al., 2021; Sorolla et al., 2022; W. Wang et al., 2021). For this reason, the CRISPR/Cas9 system has been extensively explored in this thesis. CRISPR systems mostly target foreign DNA except for type VI CRISPR/Cas13 system, which targets RNA (Abudayyeh et al., 2017).

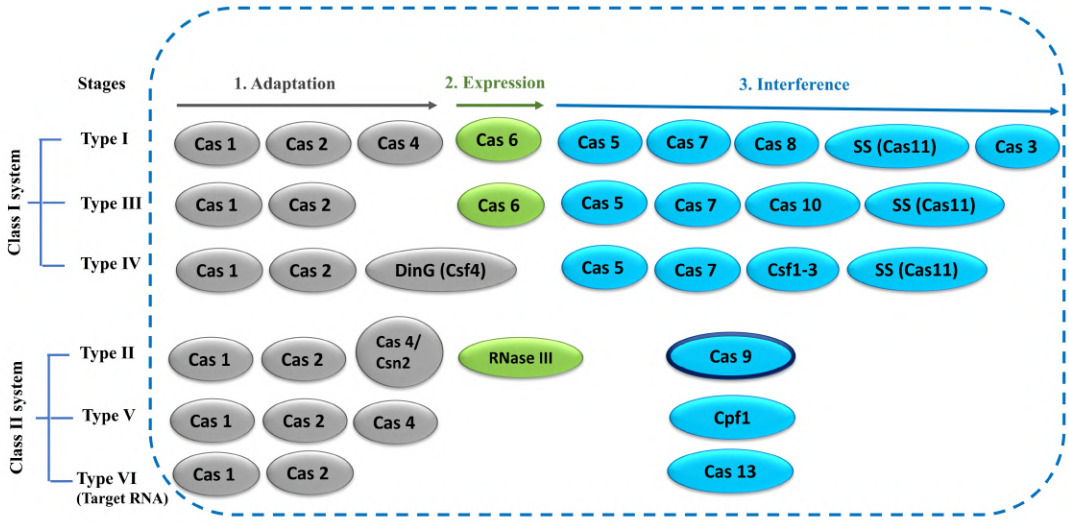


Figure 1.3. Simple overview of CRISPR-Cas classification. Cas proteins (oval) involved in various stages of CRISPR function are represented in different colours: grey (adaptation), green (expression), and blue (Interference). Cas9 protein in the type-II system is the focus of this thesis, highlighted with a dark border.

1.4. CRISPR/Cas9 system

The CRISPR/Cas9, a type IIA system, is a complex made up of single guide RNA (sgRNA) and Cas9 protein, a 160 kDa DNA endonuclease enzyme that cleaves each strand of double-stranded DNA at a precise position (Nishimasu et al., 2014). Cas1, Cas2, and Csn2 are required for the DNA acquisition step (Brouns et al., 2008; Rath et al., 2015), while RNaseIII helps in the processing of pre-crRNA to mature sgRNA (Deltcheva et al., 2011; F. Jiang & Doudna, 2017).

1.4.1. Repurposing CRISPR/Cas9 for Genome editing

The simplicity and programmability of the CRISPR/Cas9 system have enabled its rapid adaptation from a prokaryotic adaptive immune mechanism, into a robust and versatile platform for genome editing in eukaryotic cells (Hillary & Ceasar, 2023; Jinek et al., 2012). CRISPR/Cas9 systems repurposed for genome editing purposes have their crRNA and tracrRNA fused into a single, chimeric molecule known as the single guide RNA (sgRNA) (Rhun et al., 2019). The sgRNA typically contains a 20 nucleotide spacer region that is complementary to the DNA sequence of

interest, enabling precise targeting and generating a site-specific double-stranded DNA break (Le Rhun et al., 2019; Nishimasu et al., 2014).

Upon cleavage by Cas9, the cell's intrinsic DNA repair machinery is engaged to repair the double-stranded break (DSB), resulting in targeted mutations or genome editing (Carroll, 2012; Deltcheva et al., 2011; Gong et al., 2018). In the majority of cell types, the primary repair pathway employed is non-homologous end joining (NHEJ), where the separated DNA ends are connected in the absence of a homologous guide template. Although efficient, NHEJ is intrinsically error-prone and frequently introduces small insertions or deletions (indels) at the cleavage site, frequently generating frameshift mutations that disrupt gene function and facilitate the knockout of genes (Song et al., 2021; Wahab et al., 2022). In comparison, if a donor DNA template containing correct DNA sequences, is provided, the cell may recruit the homology-directed repair (HDR) pathway, that enables precise sequence alterations such as single-nucleotide substitutions or correction of pathogenic alleles, enabling targeted genome editing (Wahab et al., 2022).

Among the different Cas9 orthologs, *Streptococcus pyogenes* Cas9 (*Sp*Cas9) stands out as the most widely used variant for genome editing purposes. *Sp*Cas9 was the first Cas9 protein to have been thoroughly characterized and adapted for eukaryotic editing, thus paving the way for standardized protocols and toolkits (Barrangou & Doudna, 2016; Jinek et al., 2012). One advantage of using *Sp*Cas9 is its comparatively simple PAM requirement (5′-NGG-3′), especially in comparison with the highly restrictive PAM requirements found in some orthologs (Nishimasu et al., 2014). *Sp*Cas9 has also demonstrated robust activity across a wide variety of organisms, ranging from bacteria and yeast to plants and mammalian cells, making it an exceptionally nuclease for genome editing (Adane & Alamnie, 2024; Frangoul et al., 2021; Guiderdoni et al., 2023; Liao et al., 2021; Sadeqi et al., 2021; Wellhausen et al., 2021). These combined factors have established *Sp*Cas9 as the most widely adopted CRISPR nuclease and a central platform for genome engineering applications.

1.4.2. Domain Architecture of *Sp*Cas9

*Sp*Cas9 contains two lobes (**Figure 1.4**), namely the Recognition/REC lobe (residues 56-718) and Nuclease/NUC lobe (residues 1-55 and 719-1368) (Nishimasu et al., 2014; Ray et al., 2019). The recognition lobe contains REC-I, REC-II, and REC-III domains responsible for non-specific

binding with nucleotides: sgRNA and dsDNA (Jiang & Doudna, 2017). The arginine-rich bridge helix serves as a linker between RuvC-I and REC domains and is crucial for initiating cleavage activity upon binding to target DNA (Nishimasu et al., 2014). *Sp*Cas9 nuclease lobes have two endonuclease domains (**Figure 1.4**), the HNH domain (residues 766-909, rich in histidine and asparagine residues) and RuvC domain (residues 1-55, 719-765, and 910-1099). HNH domain cleaves the target DNA (tDNA) strand, whereas the RuvC domain cleaves the non-target DNA (ntDNA) strand (Ray et al., 2019). Additionally, HNH domain contains two key hinge regions near its N and C terminus, namely linker L1 and L2 (**Figure 1.4a**), which creates a cross-talk between RuvC and HNH domains (Palermo et al., 2016). The PAM interacting domain (P.I., residues 1100-1368) confers PAM specificity and is therefore responsible for initiating binding to DNA (Nishimasu et al., 2014; Ray et al., 2019). Upon DNA binding, the positively charged residues present at the interface between REC and NUC lobes, particularly at the bridge helix stabilize the negatively charged sgRNA:DNA hybrid. On the other hand, positively charged residues present in the linker region (L1 and L2) between RuvC and HNH domains help to stabilize the displaced ntDNA (Jiang & Doudna, 2017).

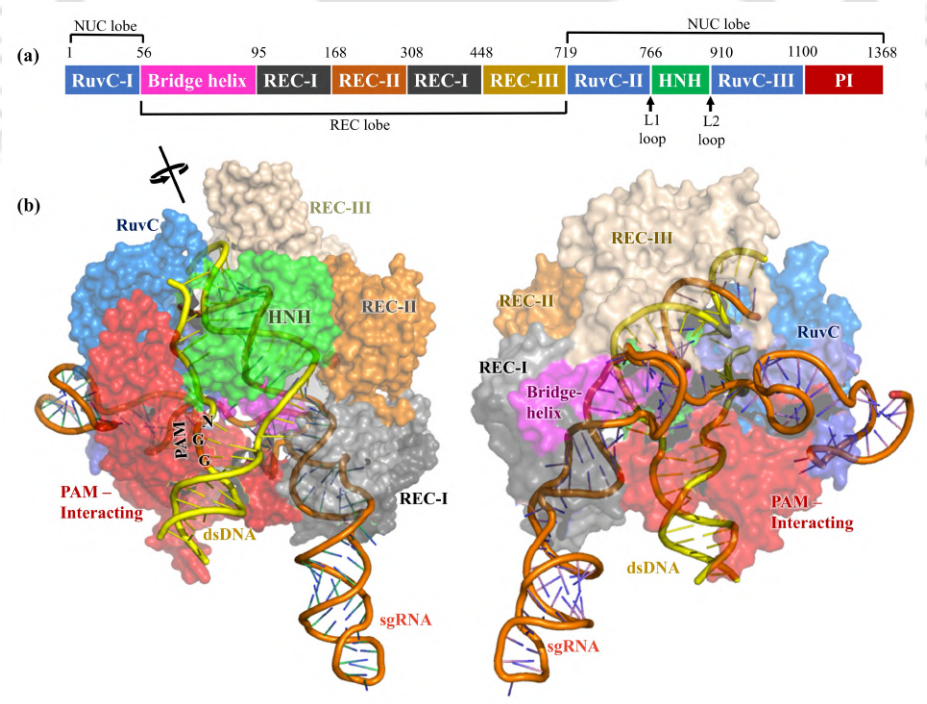


Figure 1.4. (a) Different domains of *Sp*Cas9 protein and their lengths are represented with different colours. (b) X-ray structure of *Sp*Cas9 endonuclease adopted from PDB 5F9R (resolution = 3.4 Å),

captured in a pre-cleavage state containing sgRNA (orange) and intact dsDNA (yellow). The structure on the right side is rotated by 180^0 around the axis of sgRNA. The protein is demonstrated as a transparent surface, while the dsDNA and sgRNA are visualized as cartoons.

1.4.3. Structural studies on *Sp*Cas9

X-ray structures at various stages of CRISPR pathway highlight the conformational change of *Sp*Cas9 in atomic details (**Figure 1.5**) (Anders et al., 2014; Jiang et al., 2015, 2016; Jinek et al., 2014; Nishimasu et al., 2014). Various structures along the CRISPR/Cas9 editing pathway from *Streptococcus pyogenes* have been resolved (viz., free *Sp*Cas9 (PDB 4CMQ) (Jinek et al., 2014), sgRNA bound *Sp*Cas9 (PDB 4ZT0) (Jiang et al., 2015), *Sp*Cas9 bound to tDNA and incomplete ntDNA containing PAM sequence (PDB 4UN3) (Anders et al., 2014), *Sp*Cas9 bound to both tDNA and complete ntDNA (PDB 5F9R) (Jiang et al., 2016) representing the precatalytic state of *Sp*Cas9. The apo (Jinek et al., 2014) and sgRNA bound (Jiang et al., 2015) *Sp*Cas9 structures (PDB 4CMQ and PDB 4ZT0; **Figure 1.5 (a, b)**) were resolved by Jinek et al. in 2014 and Jiang et al. in 2015 at a resolution of 3.09 Å and 2.9 Å respectively. A major rearrangement of helical REC domains of *Sp*Cas9 upon sgRNA binding was evident (**Figure 1.5 (a, b)**), with a large ~65 Å shift of REC-III domain for accommodating sgRNA (Jiang et al., 2015). Binding of tDNA and PAM containing incomplete ntDNA strand to the *Sp*Cas9:sgRNA complex (PDB 4UN3, **Figure 1.5 c**) (Anders et al., 2014) further shifts the REC-II domain in the outward directions (**Figure 1.5 (b, c)**). PAM recognition by *Sp*Cas9:sgRNA results in the melting of the foreign DNA upstream to the PAM sequence resulting in DNA:RNA hybrid formation (Anders et al., 2014; Sternberg et al., 2014). X-ray structure of *Sp*Cas9:sgRNA:tDNA complex (PDB 4OO8) (Nishimasu et al., 2014) and *Sp*Cas9:sgRNA:tDNA:incomplete ntDNA (PDB 4UN3) representing initial DNA binding state (Anders et al., 2014) complex reveal important *Sp*Cas9:DNA interactions with a noticeable conformational change of REC-II domain resulting in tDNA accommodation. The binding of a complete non-target strand is required for positioning the HNH domain close to the tDNA cleavage site (PDB 5F9R; **Figure 1.5 d**) (Jiang et al., 2016). However, the reported X-ray structure (PDB 5F9R) by Jiang et al. in 2016 **(Figure 1.4, 1.5 d)** was obtained in the absence of Mg^{2+}. Cryo-EM structures of *Sp*Cas9:sgRNA:dsDNA in the presence of Mg^{2+} resolved three key states

(Pre-catalytic: PDB 6O0Z, Post-catalytic: 6O0Y, and Product: 6O0X) (X. Zhu et al., 2019) and highlighted the importance of Mg^{2+} in stabilizing the catalytic residues of *Sp*Cas9 around the scissile phosphate and facilitate tDNA cleavage (Zhu et al., 2019). However, the location of Mg^{2+} in the catalytic pocket was unresolved in the cryo-EM structures due to poor resolution (Zhu et al., 2019). Binding of Mg^{2+} and complete ntDNA, induces a rotation of the HNH domain that brings catalytic H840 residue closer to the cleavage site (Jiang et al., 2016; Zhu et al., 2019). The significance of large rotation of HNH of *Sp*Cas9 in response to ntDNA binding has already been reported in several experimental and computational studies (Jiang et al., 2016; Palermo et al., 2017, 2016).

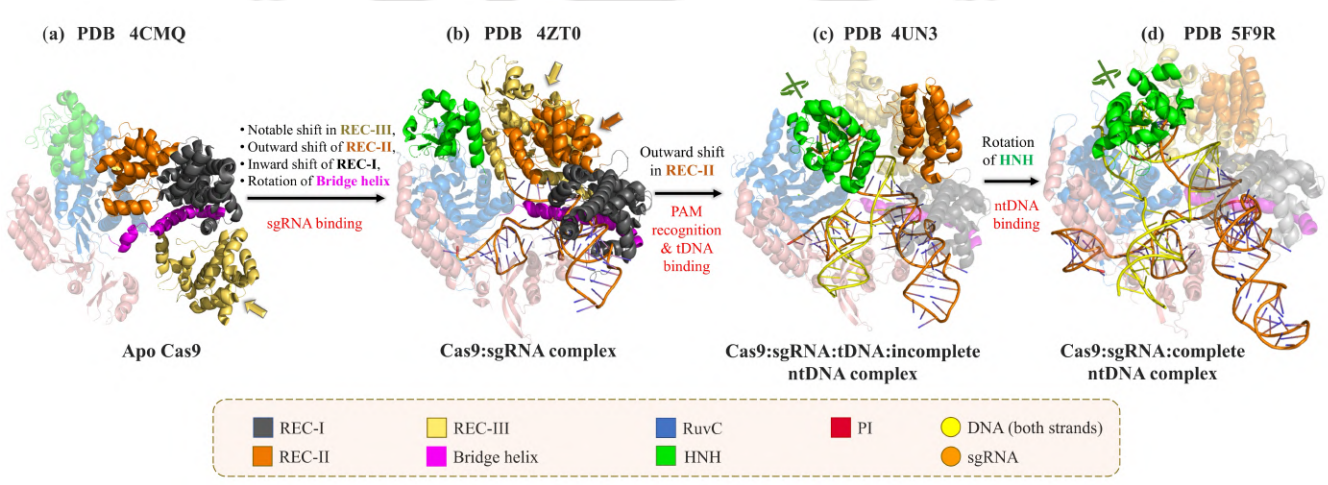


Figure 1.5. Comparison of X-ray crystal structures of Cas9 in different states (a) apo *Sp*Cas9 (PDB ID 4CMQ), (b) *Sp*Cas9:sgRNA complex (PDB ID 4ZT0), (c) *Sp*Cas9:sgRNA:tDNA complex with PAM sequence of non-target strand (PDB ID 4UN3), (d) *Sp*Cas9:sgRNA:dsDNA complex (PDB ID 5F9R). The domains showing major conformational changes between two consecutive states are highlighted as opaque cartoons (also indicated by coloured arrows), while the domains having similar structures between two consecutive stages of Cas9 are represented as transparent cartoons. The target and non-target strands of DNA are denoted as tDNA and ntDNA, respectively.

The post-catalytic state of the *Sp*Cas9 complexed with sgRNA and cleaved target DNA (PDB 7Z4J) has been structurally resolved through cryo-electron microscopy (Pacesa et al., 2022), offering a detailed snapshot of the enzyme in its post-cleavage conformation along with the

resolved position of Mg^{2+} ions in RuvC and HNH domains. Additionally, recent cryo-EM structures (PDB IDs: 9EAK, 9EAL, 9ED9, 9EDA, 9EDB) have elucidated multi-turnover states of *Sp*Cas9, revealing that post-cleavage retention of the PAM-proximal DNA product sterically hinders re-targeting and subsequent catalytic cycles (Kiernan & Taylor, 2025). These structures demonstrate that while the PAM-distal strand dissociates, persistent binding of the proximal fragment prevents the HNH domain from resetting, thereby limiting Cas9's turnover efficiency (Kiernan & Taylor, 2025).

Moreover, Bravo et al. in 2022 elucidated three Cryo-EM structures (PDB 7S4U, 7S4V, and 7S4X) depicting different intermediate stages of off-target cleavage (Bravo et al., 2022). These structures would certainly help to explore the detailed mechanism of off-target effects in the future. All the above-mentioned structural studies thus helped to understand the mechanism of action of *Sp*Cas9 in the interference step by capturing various states along the dsDNA cleavage pathway in atomic details (Anders et al., 2014; Jiang et al., 2015, 2016; Jinek et al., 2014; Nishimasu et al., 2014; Zhu et al., 2019).

1.5. PAM recognition in *Sp*Cas9 systems

The sgRNA-bound *Sp*Cas9 endonuclease looks for the target double-stranded DNA by scanning a short trinucleotide "Protospacer Adjacent Motif" (PAM; 5'-NGG-3', where N = A/T/G/C; **Figure 1.6**) that must present in the ntDNA upstream to the protospacer region (Gong et al., 2018). *Sp*Cas9:sgRNA complex targets the DNA in two steps. Initially, *Sp*Cas9:sgRNA transiently binds to DNA in a sequence-independent manner at multiple locations via random collisions for scanning of the PAM sequence (Singh et al., 2016; Sternberg et al., 2014). However, *Sp*Cas9 rapidly dissociates from non-PAM sites and upon PAM detection, it checks tDNA for complementarity with sgRNA for heteroduplex formation (residue 1-20) (Singh et al., 2016; Sternberg et al., 2014).

In *Streptococcus pyogenes* Cas9 (*Sp*Cas9), the PAM (5'-NGG-3') is recognized primarily through interactions mediated by the PAM-interacting (PI) domain within the nuclease (NUC) lobe (Anders et al., 2014; Nishimasu et al., 2014). The PI domain is positioned to interact with the

major groove of the double-stranded DNA adjacent to the target sequence. The recognition process involves residues R1333 and R1335 in the PI domain to form four direct hydrogen bonds with the guanine bases at positions 2 and 3 of the PAM sequence, respectively (G2, G3 of NGG PAM, **Figure 1.6b**), establishing a direct and base-specific recognition mechanism (Anders et al., 2014; Nishimasu et al., 2014). Arginine residues exhibit a well-established intrinsic affinity for guanine bases, owing to the guanidinium group's ability to form highly favorable bidentate hydrogen bonds with the O6 and N7 positions on the guanine Hoogsteen edge (Helene, 1977; Hossain et al., 2023, 2025). These interactions are critical for specific recognition of the NGG sequence, while rejecting the non-canonical PAM sequences.

PAM-Arginine interactions (R1333 and R1335) create a sharp kink, (Anders et al., 2014; Jiang & Doudna, 2017) which locally unwinds the DNA immediately upstream of PAM sequences in a unidirectional manner (Gong et al., 2018; Singh et al., 2016). Interestingly, this unwinding occurs without the requirement of ATP-dependent helicases (Anders et al., 2014; Jiang & Doudna, 2017). Post unwinding, a phosphate lock loop consisting of K1107, E1108, and S1109 residues stabilize the phosphate group in tDNA (+1 phosphate:1 nt upstream to PAM, **Figure 1.6**) (Anders et al., 2014; Jiang & Doudna, 2017). This rotates the +1 phosphate group and correctly orients the tDNA for base-pairing with the 20 nucleotides of sgRNA (RNA:DNA duplex) (Anders et al., 2014). RNA:DNA duplex formation results in displacement of ntDNA, leading to the formation of an R-loop structure (DNA:RNA hybrid and a displaced non-target DNA strand; **Figure 1.6**) (Palermo et al., 2017). Structural studies have shown that Cas9 induces a ~30° bend in the DNA helix to facilitate R-loop formation, which is stabilized by extensive protein-nucleic acid interactions within the REC and NUC lobes (Jiang et al., 2016; Zeng et al., 2018). The RNA:DNA duplex region are buried deep inside the positively charged groove between recognition and nuclease lobes (Nishimasu et al., 2014), thus stabilized by non-specific interactions. The unwound ntDNA strand is placed between RuvC and HNH domains of *Sp*Cas9 via electrostatic interactions of negatively charged DNA and positively charged amino acids. Interestingly, PAM complementary sequence in the tDNA does not form specific interactions with *Sp*Cas9 (Nishimasu et al., 2014).

The binding of RNA:DNA duplex and ntDNA to Cas9 trigger a conformational shift in the *Sp*Cas9 protein bringing two nuclease domains (RuvC and HNH) in contact with the DNA and introducing

a site-specific dsDNA break in the DNA (Palermo et al., 2017). Upon successful sgRNA:tDNA base-pairing, the nuclease domains in the presence of Mg^{+2} make blunt ended dsDNA cut after the 3rd nucleotide base upstream of the PAM (**Figure 1.6**) (Gong et al., 2018; Jinek et al., 2012).

As the initial checkpoint for DNA binding to *Sp*Cas9, accurate PAM recognition is fundamental to CRISPR/Cas9-based specificity (Anders et al., 2014; Nishimasu et al., 2014). The requirement for PAM allows *Sp*Cas9 to distinguish self from non-self-DNA, and prevents unwarranted cleavage at genomic sites lacking this motif (Anders et al., 2014; Nishimasu et al., 2014).

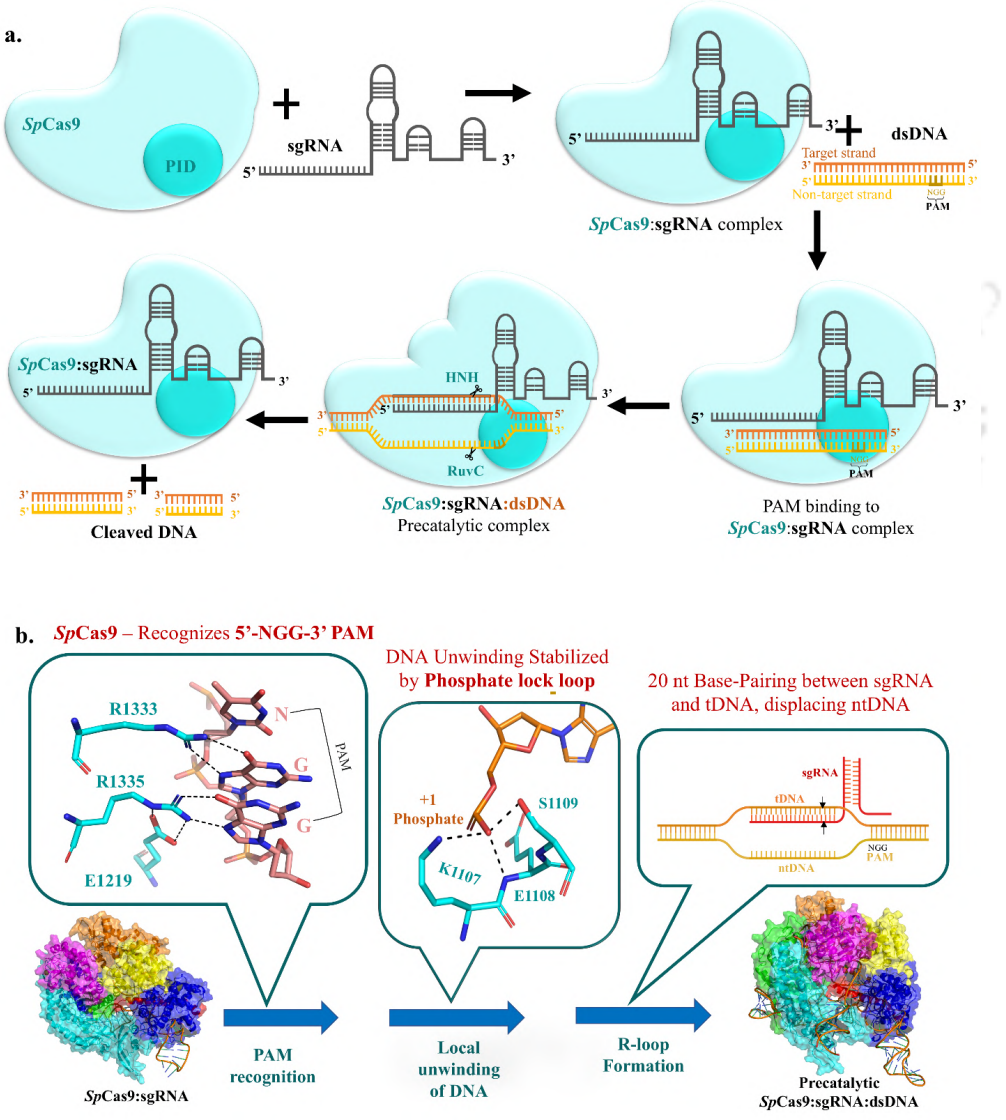


Figure 1.6. (a) Schematic representation of the mechanism of *Sp*Cas9 based target DNA cleavage. *Sp*Cas9 forms a ribonucleoprotein complex with sgRNA, which scans double-stranded DNA (dsDNA) for 5'-

NGG-3' PAM sequence. Upon PAM recognition, local DNA unwinding enables sgRNA hybridization with the target strand, forming an R-loop and displacing the non-target strand. *Sp*Cas9 then induces a double-strand break (DSB) at the target site. (b) Mechanism and consequences of 5'-NGG-3' PAM recognition by *Sp*Cas9:sgRNA complex. *Sp*Cas9 identifies the 5′-NGG-3′ protospacer adjacent motif (PAM) via key residues R1333, R1335. The PAM recognition triggers local unwinding of DNA, which is stabilized by the phosphate lock loop (K1107, E1108, S1109), anchoring the +1 phosphate. R-loop formation initiates as the sgRNA base-pairs with the target DNA (tDNA), displacing the non-target strand (ntDNA), resulting in a precatalytic *Sp*Cas9:sgRNA:dsDNA complex.

1.5.1. MD simulation-based studies on *Sp*Cas9 PAM recognition

Molecular Dynamics (MD) simulation is a popular method for investigating protein dynamics under physiological conditions at an atomic scale (Karplus & McCammon, 2002). By simulating the temporal evolution of proteins, nucleic acids, and their complexes, MD provides valuable insights into conformational flexibility, molecular motions, and interactions that are often inaccessible to static structural approaches such as X-ray crystallography or cryo-EM (Karplus & McCammon, 2002). In particular, MD enables the exploration of mechanistic pathways, conformational transitions, and energetic landscapes underlying biomolecular function, thereby bridging the gap between structural data and dynamic behavior (Karplus & McCammon, 2002).

1.5.1.1. High flexibility of PAM Interacting (PI) domain

Recognition of the PAM motif is the first step in Cas9:DNA binding. Palermo et al. in 2016 employed all-atom MD simulations to investigate the conformational dynamics of *Sp*Cas9, revealing high flexibility of PI domain (Palermo et al., 2016). Specifically, the arginine residues R1333 and R1335, which directly engage in PAM recognition, exhibit high flexibility, which aid Cas9 in searching for PAM sequences in the DNA. Moreover, the phosphate lock loop (residues K1107–S1109), which stabilizes the DNA backbone during unwinding, also displays high flexibility, which support strand separation for R-loop formation (Palermo et al., 2016).

1.5.1.2. Understanding the allosteric role of PAM sequences

The following year, Palermo et al. showed that PAM acts as an allosteric effector, which triggers interdependent conformational dynamics of HNH and RuvC catalytic domains crucial for dsDNA

cleavage (Palermo et al., 2017). Different conformations are adopted when *Sp*Cas9 interacts with PAM-containing and PAM-less DNA. Principal component analysis showed that the presence of the PAM sequence strengthens the correlation between HNH and RuvC domains (Palermo et al., 2017). PAM binding was observed to induce an "open-to-close" conformational transition in Cas9, which is crucial for nucleotide binding (Palermo et al., 2017). Residues Q771 and E584 were identified to make electrostatic interactions with K775 and R905 residues, respectively, which act as essential edges in the allosteric pathway connecting HNH and RuvC via L1 and L2 linkers (**Figure 1.7a**). PAM binding transduced signals through L1 and L2 regions. Therefore, mutations of charged residues in the L1 loops (viz., K772, T770) and important nodes in the allosteric network (viz., Q771, E584, K775, and R905) were predicted to alter enzymatic specificity (Palermo et al., 2017). In similar thoughts, Slaymaker et al. demonstrated that mutating charged residues could enhance specificity (particularly K775A mutation) and reduce the off-target effect in Cas9 by altering the allosteric signaling (Slaymaker et al., 2016). The PAM-mediated allostery mediates the essential cross-talk between RuvC and HNH domain, and plays a role in HNH activation (Palermo et al., 2016). HNH domain has been reported to undergo significant conformational changes, including a large rotation by ~180° upon ntDNA binding (**Figure 1.7b**) (Palermo, Miao, et al., 2017). dsDNA binding causes L1 and L2 domains to undergo remarkable folding and unfolding to mediate reorientation of HNH domain (Dagdas et al., 2017; Sternberg et al., 2015; Zuo & Liu, 2020), thus making it catalytically active. This highlights the significance of L1 and L2 linkers as allosteric transducers in HNH activation and mediating cross-talks between RuvC and HNH domains.

1.5.1.3. Role of D1135E mutation

Kleinstiver et al., in 2015, showed that D1135E mutation in *Sp*Cas9 improved specificity and reduced off-target effects (Kleinstiver et al., 2015). The molecular mechanism behind how D1135E mutation increases specificity for PAM recognition was explored by Kang et al. in 2022 with the help of MD simulations and free energy calculations (Kang et al., 2022). Both wildtype (W.T.) and D1135E variant were reported to maintain similar selectivity towards NGG PAM sequence (Kang et al., 2022). However, D1135E variant was demonstrated to increase discrimination against NAG PAM sequence when compared with W.T. The discrimination between NGG and NAG sequence occurred due to the lesser number of hydrogen bonding

between PAM nucleotides and R1333/R1335 residues (Kang et al., 2022). It was observed that D1135E mutation leads to a further decrease in these hydrogen bonding and thus increases the discrimination between NAG and NGG PAM sequence (Kang et al., 2022). Moreover, D1135E mutation increases the stringency of PAM recognition by breaking non-specific interactions, making DNA more reliant to base-specific interactions (Kang et al., 2022).

1.5.1.4. Exploration into the differential flexibility of R1333 and R1335 residues

A recent study has emphasized on the role of the differential flexibility of two arginine residues in PAM specificity. They anticipated that R1335 adopts in a more rigid conformation stabilized by salt bridge interactions with E1219 (**Figure 1.7c**), which might confer higher specificity by R1335 compared to more flexible R1333 residues (Hossain et al., 2025). In *Sp*Cas9:TGG complex, R1335 is sandwiched between G3 and E1219 residue creating a low entropy rigid conformation of R1335, which is hypothesized to act as a molecular "lock," ensuring precise recognition of the NGG motif (Hossain et al., 2025). In contrast, R1333 exhibits greater conformational flexibility, allowing it to sample through multiple different conformations of side chains for accommodating non-canonical bases. They further proposed that increasing the flexibility of R1335 by disrupting its stabilizing salt bridge with E1219 will increase the rotamerization of R1335 side chain allowing them to accommodate non-canonical PAM bases as shown in the xCas9 variant, which showed broadened PAM readability (Hossain et al., 2025).

Collectively, MD simulation-based studies have illuminated the intricate molecular mechanism underlying *Sp*Cas9's PAM recognition. From revealing the dynamic flexibility of key residues such as R1333, R1335, to uncovering the allosteric signaling pathways that coordinate HNH and RuvC domain activation, these investigations have deepened our mechanistic understanding of *Sp*Cas9 function. Mutational analyses, particularly involving D1135E, further underscore how alteration in residue interactions due to protein mutation can alter *Sp*Cas9:PAM specificity. Altogether, MD simulations serve as a powerful tool to bridge static structural snapshots with dynamic functional landscapes, guiding the design of next-generation genome editing platforms with improved specificity and versatility.

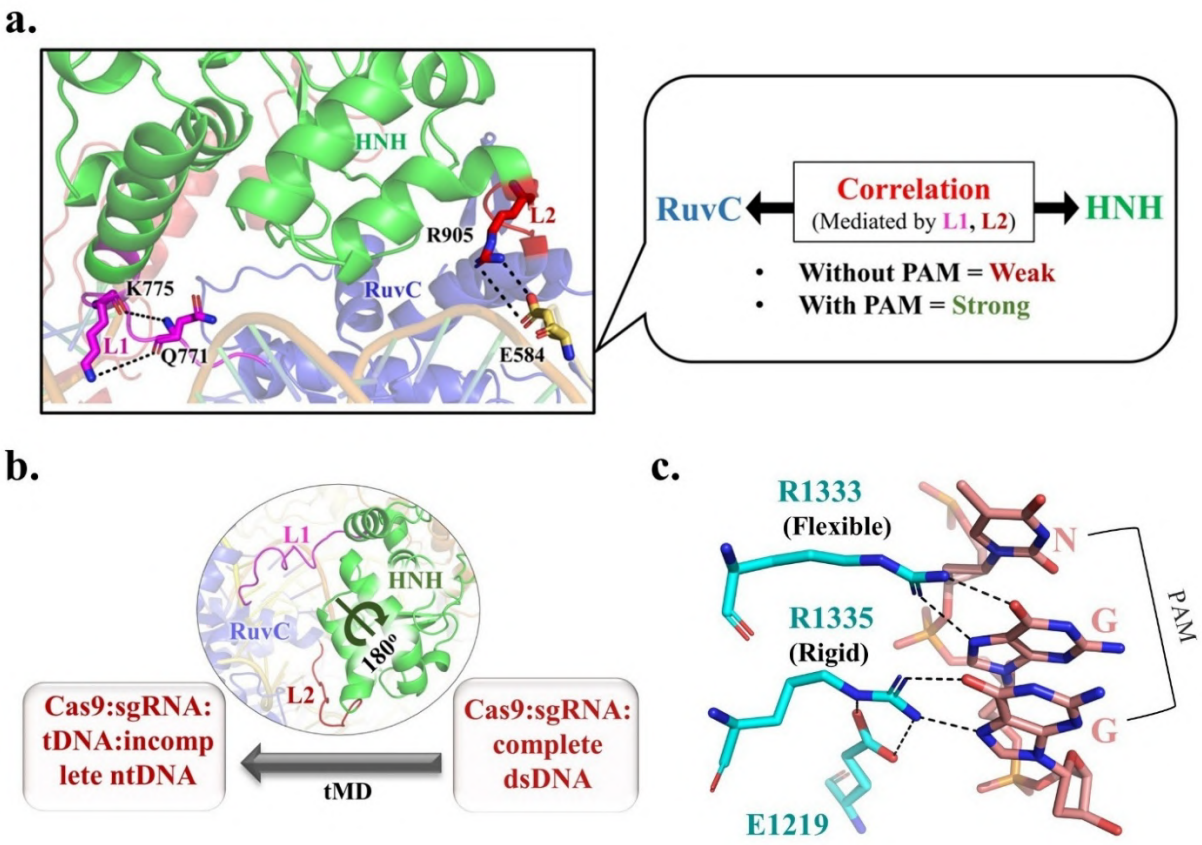


Figure 1.7. Pictorial representation highlighting the important findings from molecular dynamics studies. (a) PAM facilitates allosteric signaling (K775-Q771 and R905-E584 interactions) and establishes the correlation between HNH and RuvC domains involving L1 and L2 loops. (b) Targeted molecular dynamics (tMD) simulations of *Sp*Cas9:sgRNA:dsDNA → *Sp*Cas9:sgRNA:tDNA:incomplete-ntDNA showing rotation of HNH domains. (c) Zoomed in view of PAM binding pocket showing more rigid conformation of R1335 compared to flexible R1333.

1.5.2. Expanding *Sp*Cas9 PAM readability

Stringent PAM recognition by the CRISPR/Cas9 immune system allows bacteria to distinguish between foreign and self-DNA, but the same limits its use to genome editing applications by restricting the targetable sequences to genomic loci immediately adjacent to this trinucleotide motif. (Jinek et al., 2012; Nishimasu et al., 2014). While this stringent PAM requirement ensures high fidelity in target recognition and prevents off-target cleavage within the host CRISPR array,

it also significantly limits the number of accessible genomic sites for editing, particularly in AT-rich regions or non-coding regulatory elements lacking NGG motifs (Guo et al., 2019; Kleinstiver et al., 2015). Therefore, the expansion of PAM readability of *Sp*Cas9 is of great academic interest for technological advancement to enhance the versatility of CRISPR/Cas9 applications. *Sp*Cas9 mutations are known to affect the PAM readability (Guo et al., 2019; Kleinstiver et al., 2015). Engineering *Sp*Cas9 variants with relaxed PAM requirements enables targeting of previously inaccessible loci, facilitating broader genome coverage (Kleinstiver et al., 2015; Nishimasu et al., 2018; Ren et al., 2019; Walton et al., 2020).

This expansion is particularly valuable in therapeutic applications, where disease-associated mutations may reside in PAM-deficient regions, and precise correction is required without introducing DNA sequences. Efforts to engineer *Sp*Cas9 variants with expanded PAM readability have primarily been achieved through random mutagenesis (Kleinstiver et al., 2015; Nishimasu et al., 2018; Ren et al., 2019; Walton et al., 2020) or directed evolution (Hu et al., 2018), leading to the development of several engineered *Sp*Cas9 variants such as VQR (Kleinstiver et al., 2015), EQR (Kleinstiver et al., 2015), VRER (Kleinstiver et al., 2015), xCas9 (Hu et al., 2018), Cas9-NG (Nishimasu et al., 2018; Ren et al., 2019), SpG (Walton et al., 2020), SpRY (Walton et al., 2020), as enlisted in **Table 1.1**, which enables targeting of previously inaccessible loci. The experimental works of Kleinstiver et al. showed that mutation in *Sp*Cas9 could expand PAM sequence recognition beyond the canonical NGG, which includes NGA and NGC sequences (Kleinstiver et al., 2015). They engineered VQR, EQR and VRER variants that exhibit cleavage activity at non-canonical PAM sequences such as NGA, NGCG, and NGAG respectively by incorporating specific amino acid substitutions in the PAM-interacting domain (e.g., D1135V, R1335Q, T1337R). Moreover, Cas9 variants (xCas9) were developed that can recognize a broader range of PAM sequences via multiple mutations (Hu et al., 2018). Hu et al. designed xCas9 3.7 variant (contain seven-point mutations: A262T, R324L, S409I, E480K, E543D, M694I and E1219V) with relaxed PAM specificities (Hu et al., 2018). Guo et al. reported that among them a single E1219V mutation in *Sp*Cas9 expanded the PAM recognition by allowing cleavage activity for various noncanonical PAM sequences, 5'-GAT-3' or 5'-TGT-3' including the wildtype 5'-TGG-3 (Guo et al., 2019). They hypothesized that the broader PAM compatibility in the E1219V

variant of *Sp*Cas9 was attributed to the unrestricted rotamerization of R1335 residue (Guo et al., 2019).

Mutation of key R1335A residue in *Sp*Cas9 eliminates the base-specific interaction (R1335 and the third G of 5'-NGG-3'), thus expected to relax PAM stringency, but the activity was found to be severely compromised (Nishimasu et al., 2018). The activity of the R1335A mutant was shown to be partially restored by substituting the amino-acid residues (surrounding the PAM duplex) with arginine and or valine/phenylalanine (Nishimasu et al., 2018). Accordingly, the Cas9-NG variant was engineered which contain seven mutations in the PAM interacting domain (L1111R, D1135V, A1322R, G1218R, E1219F, R1335V, T1337R) and was shown to be active against 5'-NGA or NGC or NGT-3' PAM targets (Nishimasu et al., 2018; Ren et al., 2019). Additionally, Walton et al. engineered SpG by mutating six critical residues within the PID domain of SpCas9, eognize NGN PAM sequences (Walton et al., 2020). To further relax specificity at the second PAM base, R1333 of SpG was then mutated to proline, accompanied by substitutions of positively charged residues at additional positions in PI domain. This engineering yielded the SpRY variant, capable of recognizing NRN and NYN PAMs (where R and Y denotes purines and pyrimidines respectively), virtually removing the limitation of the rigid PAM requirement next to the DNA of interest (Walton et al., 2020). Structural analysis revealed that the mutations in SpRY created nonspecific electrostatic interactions with the DNA to compensate the loss of base-specific interactions (Walton et al., 2020).

Table 1.1. A list of engineered *Sp*Cas9 variants with expanded PAM readability.

Variant	***Sp*Cas9 Mutations**	**PAMs Recognized**	**References**
VQR	D1135V, R1335Q, T1337R	NGA	Kleinstiver et al., 2015
EQR	D1135E, R1335Q, T1337R	NGAG	Kleinstiver et al., 2015
VRER	D1135V, G1218R, R1335E, T1337R	NGCG	Kleinstiver et al., 2015

Cas9-NG	L1111R, D1135V, A1322R, G1218R, E1219F, R1335V, T1337R	NG (NGG/NGA/NCT/NGC)	Nishimasu et al., 2018
xCas9 3.7	A262T, R324L, S409I, E480K, E543D, M694I, E1219V	NGT, GAA, GAT	Hu et al., 2018
SpG	D1135L, S1136W, G1218K, E1219Q, R1335Q, T1337R	NGN (NGG/NGA/NCT/NGC)	Walton et al., 2020
SpRY	A61R, L1111R, D1135L, S1136W, G1218K, E1219Q, N1317R, A1322R, R1333P, R1335Q, T1337R	NRN > NYN (near-PAMless)	Walton et al., 2020

While effective, these strategies of creating engineered variants of expanded PAM readability are time-consuming, expensive, and may inadvertently compromise other aspects of Cas9 function, such as catalytic activity or specificity (Walton et al., 2020). The relaxation of PAM specificity must be balanced against potential trade-offs like slower kinetics or increased off-target activity, which necessitate careful consideration in the design of variants with broader PAM readability to reach an optimal balance between specificity and versatility (Hassan et al., 2021; Hu et al., 2018; Xue et al., 2023).

1.6. Motivation (knowledge gap)

*Sp*Cas9 relies on accurate recognition of 5'-NGG-3' PAM, which creates a constraint for genome editing, excluding a large fraction of genomic loci that lack a proximal NGG sequence (Kleinstiver et al., 2015; Nishimasu et al., 2018; Ren et al., 2019; Walton et al., 2020). Most protein engineering strategies that broaden PAM readability involve random mutagenesis and expensive high-throughput screening to create engineered *Sp*Cas9 variants, rather than a quantitative understanding of the energetic landscape governing PAM recognition. As a result, while biochemical assays can identify mutations in *Sp*Cas9 showing activity towards non-canonical PAM sequences, the molecular mechanisms underlying this altered PAM specificity remained unclear.

Additionally, a growing number of structural studies have elucidated the atomic details of engineered PAM-flexible *Sp*Cas9 variants. For example, high-resolution structures of Cas9-NG (PDB 6AI6), xCas9 (PDB 6AEG, 6AEB), and SpRY (PDB 8SPQ) have revealed reveal snapshots of Cas9:PAM interactions in different variants (Guo et al., 2019; Hibshman et al., 2024; Nishimasu et al., 2018). However, these static structures do not explain the role of water molecules, local environments, residue flexibility, DNA dynamics, or the energetic penalties that govern acceptance or rejection of a PAM.

As a result, the following conceptual gap remained: (a) the missing link between structures and thermodynamics; (b) the absence of a direct, quantitative framework connecting static structural and biochemical studies of wild-type *Sp*Cas9 and engineered Cas9 variants; (c) limited understanding of the structure-based free energy landscape explaining how specific SpCas9 mutations alter PAM binding affinity; and (d) the lack of a quantitative, molecular-level explanation of how *Sp*Cas9 discriminates between cognate and non-cognate PAMs.

Two central research questions arise from this foundation:

1. How do protein mutations influence the binding affinity between *Sp*Cas9 and various PAM sequences?
2. How does *Sp*Cas9 distinguish between canonical and non-canonical PAM sequences at the molecular level?

Addressing this knowledge gap is crucial. Doing so will not only create a mechanistic link between structure, thermodynamics, and biochemical studies, but will also enable the rational design of new Cas9 mutants with altered nucleotide selectivity. MD simulations are known to complement experiments and provide important insights into the molecular mechanism, structure and thermodynamics, which would help to design new mutants with altered PAM readability (Arantes et al., 2023; Casalino et al., 2020; Hossain et al., 2025; Kang et al., 2022; Mitchell et al., 2020; Palermo, 2019; Palermo et al., 2016, 2017; Palermo, Miao, et al., 2017; Ray & Felice, 2020; Ricci et al., 2019; J. Wang et al., 2023; Zuo & Liu, 2016, 2017).

This thesis seeks to systematically bridge existing gaps (structure, thermodynamics and biochemical studies) leveraging classical molecular dynamics and alchemical free energy

calculations to quantify the thermodynamic and structural basis of PAM recognition in both *Sp*Cas9 and its engineered variants, and correlating these findings directly with biochemical data. The free energy differences of DNA binding in response to protein mutation (**Chapters 2 and 3**) and protein binding in response to DNA mutations (**Chapter 4**) are extensively computed and a link between the estimated free energy differences and the structures were established in this thesis. By understanding the molecular determinants of nucleotide selectivity, this research lays the groundwork for new hypothesis, which would create mechanistic foundation for the design of new Cas9 variants that combine broadened PAM compatibility with optimal specificity, thus increasing the versatility of CRISPR/Cas9 based genome editing.

1.7. Objectives

Based on the above-mentioned knowledge gap and motivation, the objectives of this thesis are as follows:

1. Effect of single *Sp*Cas9 mutation (E1219V) on the energetics of PAM recognition.

2. Effect of multiple amino-acid mutations (L1111R, D1135V, A1322R, G1218R, E1219F, R1335V, T1337R) on the energetics of PAM recognition.

3. How does *Sp*Cas9 recognize the cognate DNA and reject the non-cognate analogue differing in the PAM sequence?

1.8. Methodology

This thesis employs classical Molecular Dynamics (MD) simulations using the X-ray structure of precatalytic *Sp*Cas9:sgRNA;dsDNA (PDB 5F9R) as a template to investigate *Sp*Cas9:nucleotide interactions, with a particular focus on protospacer adjacent motif (PAM) recognition. This structure represents *Sp*Cas9 in a catalytically competent but pre-cleavage state, bound to complete strands of the target DNA and a fully assembled sgRNA, thereby capturing the molecular

interactions that underlie target recognition and positioning prior to nuclease activation (Jiang et al., 2016). This configuration provides atomic-level insight into how the PAM-interacting (PI) domain engages the canonical 5′-NGG-3′ PAM sequence, with key arginine residues (R1333 and R1335) forming direct hydrogen bonds with guanine bases, thereby stabilizing the DNA and initiating local unwinding (Jiang et al., 2016). The structure also reveals the positioning of the HNH and RuvC nuclease domains relative to the R-loop, offering a complete view of the conformational landscape after R-loop formation (Jiang et al., 2016). These make it an ideal template for computational modeling of PAM-dependent energetics. Available structures representing the initial recognition state (PDB 4UN3) and the pre-catalytic state (PDB 5F9R) are more or less identical around the PAM recognition site (**Figure 1.8**). In comparison, the DNA content is larger for PDB 5F9R (ntDNA = 18 residue, tDNA = 30 residue) than for PDB 4UN3 (ntDNA = 11 residue, tDNA = 28 residue). Therefore, 5F9R is selected as the template for the simulation studies in this thesis.

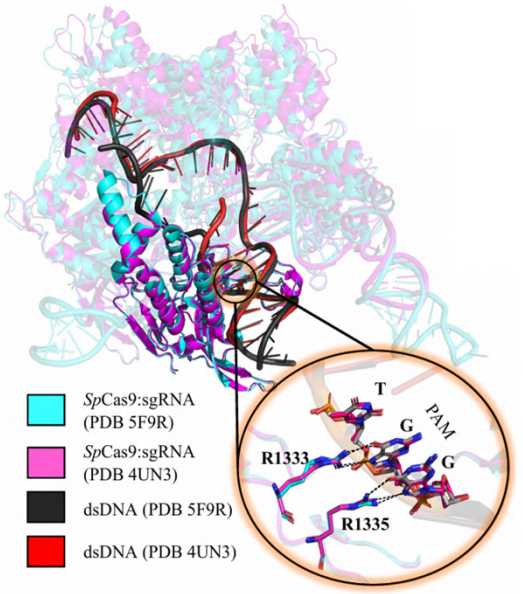


Figure 1.8. Overlay of the X-ray structures of PDB 5F9R and PDB 4UN3. The PAM interacting domains of two structures are highlighted by opaque cartoon representation (cyan for PDB 5F9R and magenta for PDB 4UN3), while other regions are rendered transparent for clarity. The bound dsDNA from PDB 5F9R and PDB 4UN3 are shown in black and red, respectively. Zoomed-in view of *Sp*Cas9:PAM interactions are shown in the circle below.

The free energy differences between canonical (5'-NGG-3') and non-canonical PAM recognition were estimated from MD trajectories by employing an appropriate thermodynamic cycle. Molecular dynamics free energy estimations serve as a powerful bridge between microscopic structure and dynamics of biomolecules and their macroscopic thermodynamic properties, most notably, the free energy, which governs biological specificity and function (Cournia & Chipot, 2024). These simulations provide atomistic insights into the energetic landscape of biomolecular recognition, offering molecular-level insights that complements experimental observations.

This chapter discusses the general theoretical frameworks and principles underlying classical MD simulations and free energy differences estimations, while the methodological details of a specific biological problem are addressed in the relevant chapters (2-4), where the application of these techniques to the *Sp*Cas9:PAM recognition is discussed in depth.

1.8.1. Principles of classical Molecular Dynamics (MD) simulations

The fundamental principle of Molecular Dynamics (MD) simulation is to model the motion of atoms by applying the laws of classical mechanics. Each atom is treated as a particle that experiences forces from neighbouring atoms, described by a mathematical potential called a force field. The simulation updates positions and velocities step by step, producing a trajectory that reveals how proteins and nucleic acids fluctuate, interact, and adapt under near-physiological conditions.

1.8.1.1. The MD Simulation Cycle

The MD cycle (**Figure 1.9**) refers to the iterative sequence of steps performed at every simulation timestep to update the atomic positions and velocities based on the forces derived from a force field. The biomolecular system used for MD simulation is defined by the initial coordinates $\{\vec{r}_i\}(0)$ of each atoms (i) and initial velocities $\{\vec{v}_i\}(0)$. The initial atomic coordinates of the biomolecule are usually obtained from experimentally determined structures (X-ray, NMR, CryoEM) or theoretically modelled systems. In this thesis, the initial structures are derived from the x-ray structures of precatalytic *Sp*Cas9:sgRNA:dsDNA complex retrieved from PDB 5F9R (Jiang et al., 2016). The initial velocities are randomly generated, typically drawn from a Maxwell–Boltzmann distribution corresponding to the desired simulation temperature.

Next, the potential energy $U(r)$ as a function of atomic positions of the biomolecule is defined by a force field, which contains set of equations and parameters that represents all the atomic interaction. From the computed potential energy U(r), the force can be described as the negative gradient of the potential energy ($U(r)$) (**equation 1.1**).

$$\boldsymbol{F}_i = -\frac{\partial U(\boldsymbol{r})}{\partial \boldsymbol{r}_i} = -\nabla_{\mathbf{r}_i} U(\mathrm{r_i}) \tag{1.1}$$

Here, $\boldsymbol{r}_i$ denotes the position vector of particle i in three-dimensional space, $\boldsymbol{r}_i = r_1,\ldots r_n$, $\boldsymbol{F}_i$ represents the force acting on the particle i, $U(r)$ denotes the potential energy function of the position of all the atoms, and $\nabla_{\mathbf{r}_i}$ is the gradient $(\frac{\partial U(\mathbf{r})}{\partial \mathbf{r}_i})$ of the potential energy with respect to the position of particle i.

Also, according to the newton's second law of motion (**equation 1.2)**,

$$\boldsymbol{F}_i = m_i \boldsymbol{a}_i = m_i \ddot{\boldsymbol{r}}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} \tag{1.2}$$

Here, m_i and $\boldsymbol{a_i}$ represents mass and acceleration of particle *i* respectively, while $\ddot{\boldsymbol{r}}_i$ denotes the second derivative of the position ($\boldsymbol{r}$) of the particle i $(\frac{d^2 \mathbf{r}_i}{dt^2})$ with respect to time, such that acceleration $\boldsymbol{a}_i = \ddot{\boldsymbol{r}}_i$.

Combining **equation 1.1** and **equation 1.2**, we arrive at the central equation for MD simulation (**Equation 1.3**).

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -\nabla_{\mathbf{r}_i} U(\mathrm{r}) \tag{1.3}$$

The primary objective in MD simulation is to simulate the time evolution of a collection of *N* particles (such as atoms in a molecule) by numerically integrating their equations of motion. The **equation 1.3** represents the Newton's equation of motion that links the time-dependent changes in atomic coordinates to the gradient of the potential energy, thereby governing the dynamic evolution of the system. Generating MD trajectories requires solving this equation to obtain updated positions and velocities. Since exact solutions are not feasible for large systems, numerical integration algorithms are used. The most commonly used numerical integration algorithm is the Velocity Verlet algorithm (Swope et al., 1982) for solving **Equation 1.3**. Starting from known positions $\mathbf{r}_i(t)$, velocities $\mathbf{v}_i(t)$, and forces $\mathbf{F}_i(t)$ on each particle i at time t, the Velocity Verlet algorithm updates the system at the next timestep $t + \Delta t$ in two steps. First, the positions of the particles are updated using **Equation 1.4**. computes the new position $\mathbf{r}(t + \Delta t)$ by advancing the current position $\mathbf{r}(t)$ based on the current velocity $\mathbf{v}(t)$ and the acceleration (force $\mathbf{F}(t)$ divided by mass m) over the squared timestep Δt^2.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\frac{\mathbf{F}(t)}{m}\Delta t^2 \quad (1.4)$$

Next, the new updated velocity $\mathbf{v}(t + \Delta t)$ is calculated by averaging the acceleration at the current and the newly computed positions (forces $\mathbf{F}(t)$ and $\mathbf{F}(t + \Delta t)$) over the timestep Δt, as shown in **Equation 1.5**.

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{1}{2m}[\mathbf{F}(t) + \mathbf{F}(t + \Delta t)]\Delta t \quad (1.5)$$

This averaging improves accuracy and ensures smooth velocity evolution consistent with the position update. In this thesis, a timestep Δt of 2 femtoseconds (fs) is chosen for numerical integration in all molecular dynamics simulations. This timestep is enabled by constraining high-frequency bond vibrations involving hydrogen atoms, which removes the fastest vibrational modes from the system and allows stable numerical integration with a longer timestep. The chosen timestep therefore balances computational efficiency with numerical stability and accuracy, enabling reliable simulations of biomolecular dynamics over extended timescales.

Once the atomic positions and velocities are updated through numerical integration, the new coordinates are used to recalculate the forces acting on each particle, thereby initiating the next cycle of the Molecular Dynamics (MD) simulation. This iterative MD cycle is repeated is repeated millions of times (0 → dt → 2dt → 3dt →…… ndt) over the course of a simulation, where dt corresponds to timesteps. The result is a trajectory that records the atomic coordinates and velocities at successive time intervals, typically spanning nanoseconds to microseconds of simulated time. This trajectory can be used for extracting average structural, dynamical, and thermodynamic properties of biomolecular systems.

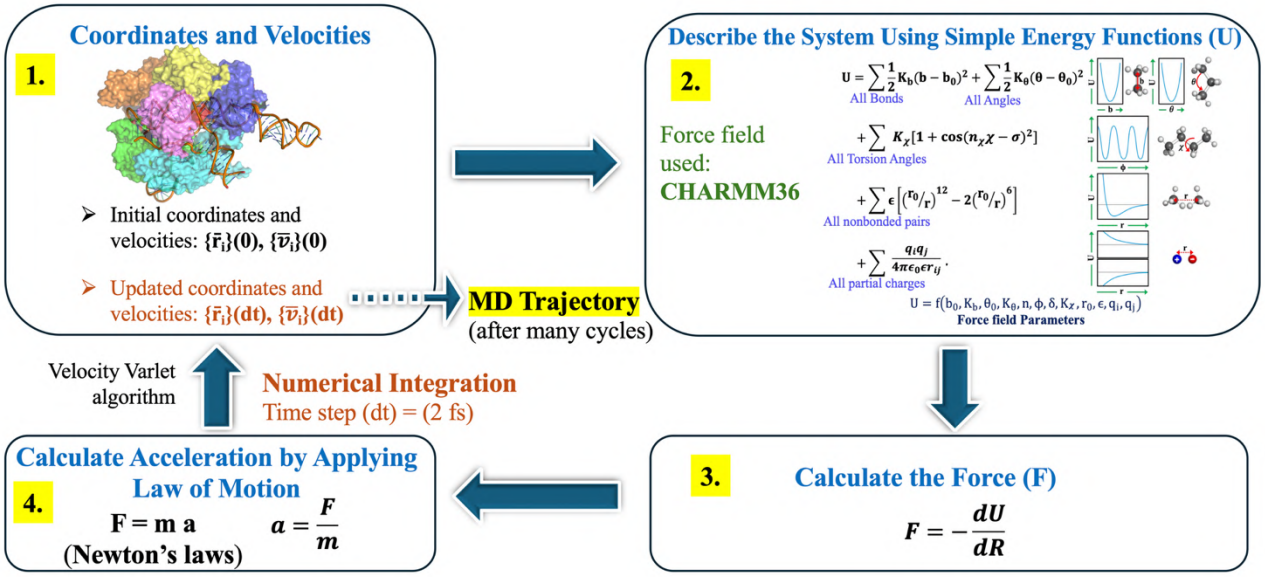


Figure 1.9. The MD cycle. The process consists of four steps: (1). Defining initial atomic coordinates and assigning initial velocities, (2). Evaluating the system's potential energy (*U(r)*) using a force filed, (3). Calculating force as a negative gradient of the potential energy with respect to its position, (4). Applying newtons law of motion to calculate acceleration followed by numerical integration to obtain updated coordinates and velocities. Repeating this cycle over millions of steps generates a trajectory that captures the motion of atoms with time enabling analysis of structural and thermodynamic properties.

1.8.1.2. Force Fields

In molecular dynamics (MD) simulations, force fields are the set of mathematical functions and associated parameters used to describe the potential energy of a molecular system as a function of atomic coordinates in it's three-dimensional structure. They define identity of a biomolecule by providing the necessary equations and parameters to describe atomic interactions, both bonded and non-bonded, allowing the computation of forces responsible for molecular motion. The potential energy function (*U(r)*) serves as the fundamental basis for calculating interatomic forces in MD simulations.

In a force-field, atoms are simplified as a sphere with a van der Waals radius and partial charge, and bonds are represented as springs with force constant k. The total potential energy *U(r)* is can

be represented as the sum of bonded potential energy ($U(r)_{\text{bonded}}$) and non-bonded potential energy $U(r)_{\text{non-bonded}}$ terms (**equation 1.6**).

$$U(r) = U(r)_{\text{bonded}} + U(r)_{\text{non-bonded}} \tag{1.6}$$

where bonded interactions include bond stretching, angle bending, dihedral torsions, improper torsions, and other auxiliary bonded terms defined within a given force field, while non-bonded interactions comprise van der Waals forces and electrostatics. Combining both bonded and non-bonded terms, a typical equation representing the potential energy function (*U(r)*) by a force-field is provided in **equation 1.7,** which allow us to estimate the potential energy landscape of a biomolecule.

$$U_{total} = \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\chi[1 + \cos(n_\chi\chi - \sigma)^2] + \sum_{\text{nonbonded pairs},ij} \left(\epsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0\epsilon r_{ij}}\right) \tag{1.7}$$

The first three terms represent bonded interactions (bond length, angles and dihedrals; **Equation 1.8, Figure 1.10**)

$$U_{bonded} = \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\chi[1 + \cos(n_\chi\chi - \sigma)^2],) \tag{1.8}$$

The first bonded term ($\sum_{\text{bonds}} K_b(b - b_0)^2$) include the energy associated with the change in bonds lengths by bond stretching, where K_b is the spring constant, b is the current bond length, and b_0 is the equilibrium bond length, and $(b - b_0)$ indicates the net displacement of the bond from equilibrium bond-length (b_0). The second term ($\sum_{\text{angles}} K_\theta(\theta - \theta_0)^2$) represents the potential energy associated with deviation of bond angle θ from the equilibrium bond angle θ_0, with K_θ as the force constant. Both theses terms are modeled using the harmonic potential, which increases the potential energies as bond length and angles deviates from its equilibrium values. The third term ($\sum_{\text{dihedrals}} K_\chi[1 + \cos(n_\chi\chi - \sigma)^2]$) defines the potential energies associated with change in dihedral (or torsional) angles, where χ is the torsional angle, K_χ is the dihedral force

constant, n_χ is the multiplicity (number of minimum points as the torsional angle is rotated by 360°), and σ is the phase shift (torsion angle where $U_{dihedral}$ will be maximum).

On the other hand, fourth and fifth terms in the potential energy function (**equation 1.7, Figure 1.10)** corresponds to non-bonded (non-covalent) interactions (van der waals and electrostatics respectively; **Equation 1.10**)

$$U_{non-bonded} = \sum_{\text{nonbonded pairs},ij} \left(\epsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 \epsilon r_{ij}} \right) \quad (1.9)$$

The fourth term ($\epsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min,ij}}{r_{ij}} \right)^{6} \right]$) represents the van der Waals or Lennard-Jones (LJ) potential, which models the interaction between neutral atom pairs *i* and *j*. It comprises of (1) short-range repulsive component proportional to r^{-12}, preventing atomic overlap by creating a steep energy barrier as atoms approach each other very closely, and (2) an attractive van der Waals component proportional to r^{-6}, which captures the long-range dispersion forces responsible for weak attraction between neutral atoms or molecules. Here, ϵ_{ij} represents the depth of the potential well indicating the strength of the interaction, and $R_{\min,ij}$ is the distance at which the potential energy reaches its minimum value These interactions are short range interactions, and thus degrades very fast over larger distances.

The fifth term in the potential energy function ($\frac{q_i q_j}{4\pi\epsilon_0 \epsilon r_{ij}}$) represents the Coulombic interaction potential, which describes the electrostatic interactions between charged atoms *i* and *j*. This interaction can be either repulsive or attractive, depending on the sign of the partial charges q_i and q_j carried by each atom. Here, r_{ij} is the distance between atoms *i* and *j*, ϵ is the relative dielectric constant of the medium and ϵ_0 is the permittivity in vacuum.

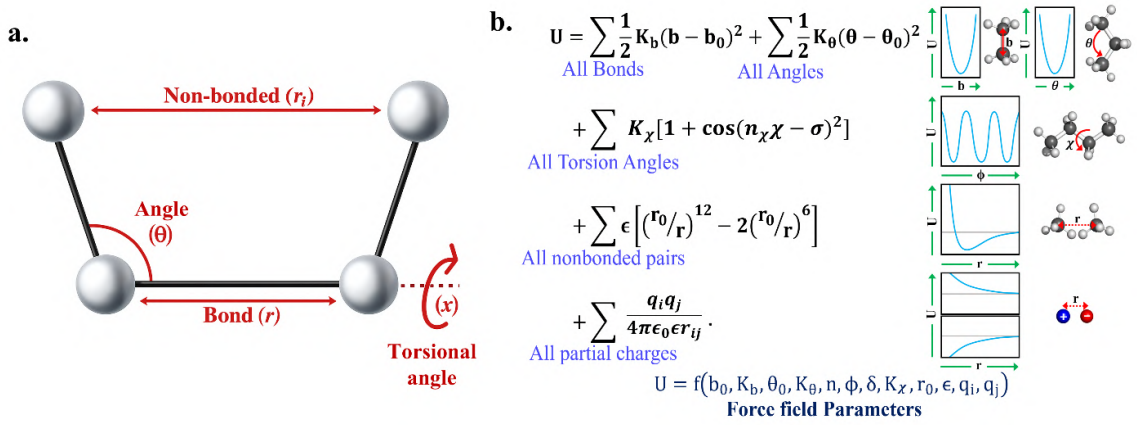


Figure 1.10. Schematic representation of molecular mechanics terms in a classical force field: bond stretching (r), angle bending (θ), torsional angle (x), and non-bonded interactions (r_i). The total potential energy function ($U(r)$) used in force fields, including harmonic terms for all bonds and angles, periodic functions for torsion angles, Lennard-Jones potentials for non-bonded pairs, and Coulombic interactions for partial charges. Representative energy profiles, force field parameters and molecular examples are shown for each interaction type.

Force-field parameters are generally derived either from experimental data or from high-level quantum chemical calculations (Fröhlking et al., 2020; Huang & Mackerell, 2013) and are kept constant within a given classical force field. Currently, a wide variety of force fields are available for simulation of biological systems such as CHARMM (Brooks et al., 1983, 2009), AMBER (Ponder & Case, 2003), GROMOS (Scott et al., 1999) and OPLS (Jorgensen et al., 1996). In this thesis, the CHARMM36 (Chemistry at HARvard Macromolecular Mechanics version 36) force field is employed, since it is a widely validated and extensively optimized designed for accurate modeling of diverse biomolecular systems, including proteins, nucleic acids, lipids, and carbohydrates (Huang & Mackerell, 2013), making it an excellent choice for studying *Sp*Cas9:sgRNA;dsDNA complex. In the CHARMM36 force field, the total potential energy includes additional bonded interaction terms beyond those explicitly written in **Equation 1.7**, such as improper dihedral and Urey–Bradley interactions, which are used to maintain planarity, chirality, and appropriate 1–3 interactions in biomolecular systems.

1.8.1.3. Solvation and water models

In this thesis, the biomolecular system is overlaid on a pre-equilibrated explicit water box (**Figure 1.11b**), keeping the biomolecule at the center, with a minimum distance of 1.2 nm between the biomolecule's surface and the box edge to ensure adequate solvation. This spacing corresponds roughly to about three layers of water molecules fully surrounding the biomolecule, which helps to replicate the solvation environment found in physiological conditions such as hydration shell formation, dynamic water-mediated interactions etc. TIP3P water model (Jorgensen et al., 1998) is used for performing MD simulations. This model has a simple a rigid, non-polarizable, three-site representation corresponding to the three atoms of the water molecule (**Figure 1.11a**), which remains the most popular choice for simulating biomolecular systems (Jorgensen et al., 1998). To maintain molecular rigidity, intramolecular bond lengths within water molecules were constrained using the ShakeH algorithm (Ryckaert et al., 1977) or LINCS algorithm (Hess, 2007; Hess et al., 1997). The solvated system was subsequently neutralized by the addition of appropriate counter-ions to ensure charge neutrality (**Figure 1.11c**).

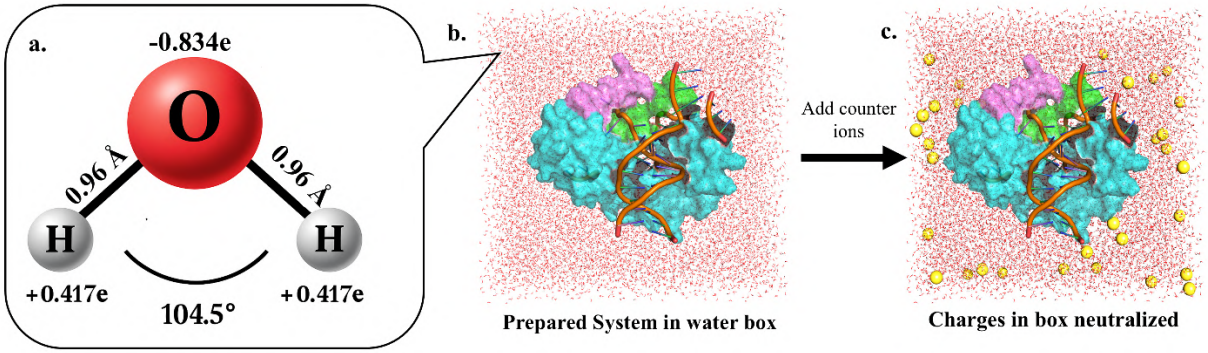


Figure 1.11. Solvation of Biomolecular System Using TIP3P Water Model. (a) Structural representation of a water molecule modeled using TIP3P parameters, showing partial charges (−0.834e on oxygen, +0.417e on hydrogen), bond lengths (0.96 Å), and bond angle (104.5°). (b) Solvated biomolecular system placed within an explicit water box. (c) Charge neutralization of the simulation system by addition of counter-ions.

1.8.1.4. Periodic Boundary conditions

In molecular dynamics (MD) simulations, systems are modeled using a finite number of particles confined within a simulation box. However, most real-world biomolecular systems exist in bulk

(macroscopic) environments (N ~ 10^{23} atoms), far larger than what is computationally feasible (less than 10^6 atoms). If boundaries of the box are treated as non-periodic walls facing vacuum, particles near the edges will experience unnatural forces and densities, leading to significant surface or boundary effects that can distort results and prevent the realistic sampling of bulk properties. Therefore, MD simulations of finite size in a box of water must be carried out in such a way that it mimics the bulk solvent. To solve this artifact, periodic boundary conditions (PBCs) are applied on the simulation box to remove surface artifacts and include bulk effect in the simulation systems (Rahman & Stillinger, 1971). Under PBCs, the simulation box (often called the "unit cell") is conceptually surrounded by infinite replicas of itself in all directions. Each unit cell is surrounded by eight neighboring cells for a two-dimensional representation of PBCs (**Figure 1.12)**. In three dimensions, a unit cell will have 26 neighboring cells. The PBCs ensure that when a particle exits one face of the box, an imaged particle seamlessly re-enters from the opposite face with the same velocity, keeping the number of particles in the central unit cell constant throughout the MD simulation. Thus, every particle is always surrounded by an environment identical to the original system, mimicking an infinite system, where no atoms feel any surface forces.

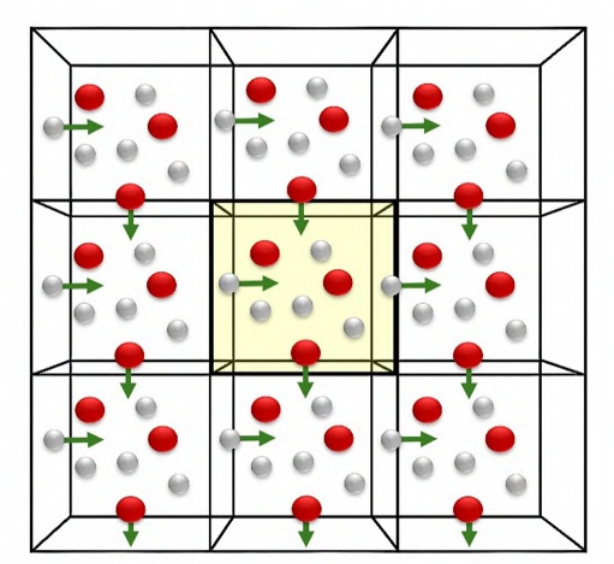


Figure 1.12. Schematic representation of Periodic boundary conditions (PBCs). The unit cell (central cell, yellow) is replicated in all directions. Red and gray spheres denote particles simulated. If a particle leaves the cell (arrows), another particle will enter from replica cell and replace the particle left in main cell.

1.8.1.5. Short-range Van der Waals Interactions

Van der Waals (vdW) interactions, which are modeled using the Lennard-Jones (LJ) potential, are weak, non-covalent forces that arise from transient fluctuations in electron density between atoms and molecules. These forces are short-range in nature and decays rapidly, decreasing in strength with distance roughly proportional to r^{-6}, where r is the distance between interacting particles. Therefore, it is a widely adopted strategy in MD simulations to apply a cut-off distance when calculating LJ potential (Frenkel & Smit, 2023). This approach significantly improves computational efficiency by ignoring interactions beyond a specified cutoff range where the forces become negligible.

However, A cutoff function abruptly truncates the LJ potential energy of non-bonded interaction, which can introduce artifacts into the simulation. To mitigate this, a switching function is often employed. This function smoothly decreases the LJ potential to zero over a small range near the cutoff distance, ensuring a gradual decay of the interactions. In our simulations, we implemented a cut-off range of 12 Å, with the switching function applied over the final 1 Å to ensure a gradual decay of LJ interactions.

1.8.1.6. Long-range Electrostatic Interactions

Electrostatic interactions, i.e. forces between charged or partially charged atoms are fundamental to the behaviour of biomolecules and play an important role in protein:nucleic acid complex stabilization. These interactions are most computationally expensive due to their long-range nature. Unlike van der Waals forces, which decay rapidly with distance, Coulomb interactions decay slowly, only inversely with distance as r^{-1}. Thus, electrostatic interactions need to be computed over long distances without truncation. In MD simulations, Particle Mesh Ewald (PME) method (Darden et al., 1993) is used to estimate the electrostatic interactions in presence of periodic boundary conditions. In this algorithm, the electrostatic potential is divided into two distinct components: a short-range part that decays rapidly and is computed in real space, and a long-range part that decays slowly and is evaluated in reciprocal (Fourier) space. Accurate treatment of the long-range interactions necessitates both charge neutrality and periodic boundary conditions within the simulation system.

1.8.1.7. Energy minimization

Experimentally determined structures such as X-ray crystallography, Cryo-EM, or NMR based structures may exhibit localized strains like steric clashes or atomic overlap arising from poor resolution. If the initial structure contains unfavourable contacts, the strains in the structures can be relieved by lowering the potential energy, until it reaches the local minima of the potential energy hypersurface. The potential energy hypersurface of a system is a function its atomic coordinates, and typically contains multiple local minima due to the high degrees of freedom associated with the biomolecules. Any strain in the molecule creates displacement from these minima, which leads to an increase in potential energy, as the system moves away from its energetically favorable state. Therefore, energy minimization is performed by systematically adjusting atomic coordinates based on the forces derived from a force field, minimizing the potential energies of the molecule. The process continues until the forces become sufficiently small, indicating that the system has reached very close to the local minima of the potential energy hypersurface. Two widely used first-order minimization algorithms: (1) steepest descents (Deift & Zhou, 1993; Meza, 2010) and (2) conjugate gradient method (Hestenes & Stiefel, 1952) are employed for energy minimization in this thesis to relieve unfavourable contacts in the initial structure before performing MD. This ensures that the system starts in a physically stable, low-energy configuration without atomic overlaps that could cause instabilities or artifacts during MD simulation.

1.8.1.8. Temperature and Pressure Control

Temperature of the simulation is related to average kinetic energy of the atoms being simulated. Controlling temperature is essential to ensure that the simulated system accurately reflects the desired thermodynamic conditions. Temperature control in MD simulations was achieved by using either the Langevin dynamics (Feller et al., 1995) or velocity rescaling (Bussi et al., 2007), which modifies the momenta of all atoms to maintain a target temperature. Pressure control was implemented via the Nose-Hoover barostat algorithm (Martyna et al., 1994) or Parrinello–Rahman barostat (Nosé & Klein, 1983; Parrinello & Rahman, 1981), where the simulation box volume is dynamically adjusted to preserve a constant pressure. While instantaneous pressure values in biomolecular systems exhibit frequent fluctuations, the time-averaged pressure across

all particles are more stable and represents the system's total pressure (Hoover et al., 1982; Nosé, 1984; Hoover, 1985). Together, these methods enable the molecular system to closely mimic experimental conditions of constant temperature and pressure.

1.8.2. MD Setup adopted in this thesis

The Molecular Dynamics (MD) setup is schematically illustrated in **Figure 1.14**. Two distinct types of initial structures were adopted as a template in this thesis for performing MD simulations.

(a) **Full System:** The complete structure of the precatalytic *Sp*Cas9 bound to sgRNA and dsDNA (containing canonical 5'-TGG-3' PAM sequence) was retrieved from Protein Data Bank (Berman et al., 2000) (PDB 5F9R, resolution = 3.4 Å) (Jiang et al., 2016).

(b) **Spherically Truncated System**: A smaller, 25 Å spherically truncated system centered at the residue of interest was extracted from the full PDB structure. A spherically truncated sphere can be made sufficiently large to include the entire *Sp*Cas9/nucleotide complex. However, the advantage of the truncated system (typically about 25-30 Å in radii) lies in the possibility of cutting out a particular region of interest, which not only reduces the computational cost but also considerably improves the convergence of free energy calculations by the exhaustive sampling of the desired phase space (Lind et al., 2019). If the goal is not to sample conformational changes away from the region of interest, but rather to obtain reliable converged free energy estimates (which certainly requires multiple independent simulations), then the truncated approach has been shown to be very useful. Truncated models are more efficient than large-scale models, due to the fact that the former avoids sampling distal larger-scale motions that require longer time scales for convergence (Lind et al., 2019). This has been shown that small truncated systems (between 20Å to 30Å radii) are advantageous for studying complex systems such as the ribosome, for which extensive free energy calculations would be impractical due to the large size. Hence, it is not surprising that simulations with spherically truncated systems have been completely dominating in the literature on binding free energy calculations, which include protein-nucleotide complexes (Allnér & Nilsson, 2011; Almlöf et al., 2007; Kumar et al., 2017; Satpati et al., 2014; Trobro & Åqvist, 2005; X. Zeng et al., 2014). To ensure the robustness and reliability of this spherically truncated system, the free energy calculations (as described in section 1.8.3) were performed using both the truncated

and full systems, allowing direct comparison and assessment of the truncated model's accuracy in capturing relevant energetics. The results were averaged over multiple independent simulations to increase sampling. Furthermore, comparison of the MD data (viz., structures, ΔΔG, etc.) with the experiments (if available, viz., X-ray, thermodynamic data, etc.) ensure accuracy.

These structures were prepared by assigning appropriate topology parameters and adding hydrogen atoms according to the expected protonation states at physiological pH, generating coordinate files for molecular dynamics simulations. The standard CHARMM36 force field was used throughout to model biomolecules (Huang & Mackerell, 2013; MacKerell et al., 1998). The parameters used for MD simulations is provided in **Table 1.2**.

The resulting structure was positioned in the center and overlaid with an explicit water box, ensuring a minimum distance of 12 Å between the biomolecule and the edges of the simulation box. The TIP3P model was used to represent the water molecules (Jorgensen et al., 1998). Appropriate counter ions were introduced (sodium (Na^+) or chloride (Cl^-) based on the charge of the system) into the solvated simulation box to neutralize the overall charge of the molecular system. This is an important step because biomolecules such as proteins and nucleic acids often carry net charges that, if left unbalanced, would create artificial electrostatic effects and unrealistic forces under periodic boundary conditions. Using monovalent ions for charge neutralization is a common protocol in MD simulations of related protein–nucleic acid complexes, where the emphasis is on binding recognition rather than catalytic activity (Cheatham & Case, 2013; Torella et al., 2010; Van Heesch et al., 2023; Yamaguchi et al., 2002). Modelling divalent ions (such as Mg^{2+}, Ca^{2+}, etc.) for charge neutralization using classical fixed-charge force field is limited by the lack of explicit polarization effects, strong site-specific coordination, and slow binding-unbinding kinetics hindering adequate sampling and convergence (Allnér et al., 2012; Li & Merz, 2013; Panteva et al., 2015). Thus, monovalent ions are generally preferred in classical MD simulations, as they provide a reasonable electrostatic screening environment without introducing such complexities (Yoo & Aksimentiev, 2012).

The resulting system was subjected to energy minimization by employing the steepest descent (Deift & Zhou, 1993; Meza, 2010) or conjugate (Hestenes & Stiefel, 1952) algorithm for up to

50,000 steps. This step relieves steric clashes and unfavorable contacts. The solvated and energy-minimized system was assigned initial velocities generated from a Maxwell–Boltzmann distribution, which was then gradually heated to desired temperature (310 K) over the course of the equilibration phase.

Sequential equilibration was performed in both the NVT (constant number, volume, and temperature) and NPT (constant number, pressure, and temperature) ensembles. Heavy atoms were harmonically restraint (Force constant = 10 kcal.mol^{-1}Å^{-2} , relative to the X-ray structure. The non-hydrogen atoms of the outer regions (between 23 Å to 25 Å) of the 25 Å truncated biomolecule were harmonically restrained throughout the MD trajectory. The restraint was gradually increased from 1.0 to 5.0 kcal.mol^{-1}Å^{-2} towards the boundary (**Figure 1.13**). Harmonic restraint (Force constant = 10 kcal.mol^{-1}Å^{-2}) for the inner 22 Å radius was applied only in the equilibration phase, which was gradually removed and made completely restraint-free during production MD. In the latter stage of equilibration, the restraints were gradually removed and in the final equilibration phase and during the production run, all the harmonic restraints were removed except the buffer restraints in the case of truncated models. The full-length systems were kept completely restraint-free in the last equilibrium stage and during production dynamics.

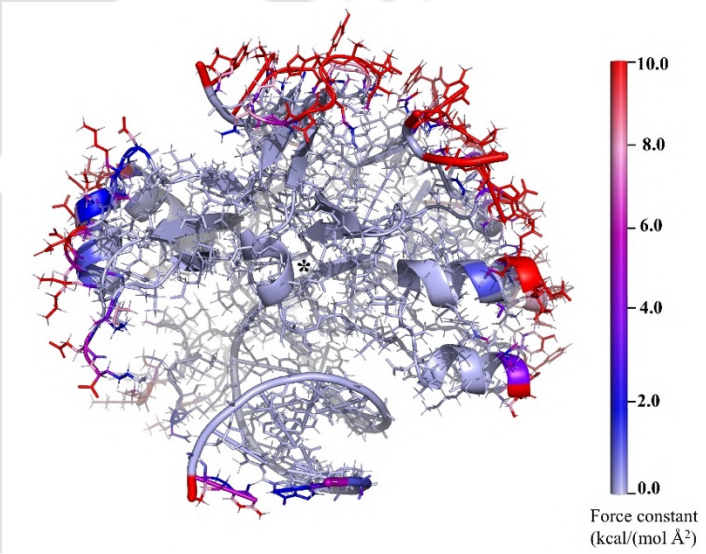


Figure 1.13. Pictorial representation of the distribution of harmonic restraints at the outer or buffer region (23-25 Å) of a 25 Å spherically truncated biomolecule. The outermost region (> 24.5 Å) has the highest harmonic restraint of force constant = 10 kcal mol^{-1} Å^{-2}, which gradually decreases to 1 kcal mol^{-1} Å^{-2} as it reaches 23 Å radius. No restraint was employed for the inner 23 Å region; thus, it is fully flexible during the production dynamics.

Following equilibration, unrestrained production dynamics were performed in the NPT ensemble. Simulations were performed at 310 K under constant pressure conditions of 1 bar Periodic boundary conditions were applied in all directions throughout the simulations, using a 2 fs integration time step. Temperature was controlled by employing either Langevin dynamics (Phillips et al., 2005) (Chapters 2 and 3) or

velocity rescaling (Bussi et al., 2007) (Chapter 4) applied to non-hydrogen atoms. The pressure was maintained using the Langevin piston (Nose Hoover method) (Feller et al., 1995; Martyna et al., 1994) (Chapters 2 and 3) or Parrinello–Rahman barostat (Nosé & Klein, 1983; Parrinello & Rahman, 1981) (Chapter 4).

Van der Waals interactions were truncated at a cutoff distance of 12 Å, while long-range electrostatics were computed using the Particle Mesh Ewald (PME) method (Darden et al., 1993) with a cutoff of 1.2 nm for short-range electrostatic interactions. The ShakeH algorithm (Phillips et al., 2005) (chapter 2 and 3) or LINCS algorithm (Hess, 2007; Hess et al., 1997) (chapter 4) was employed to restrain the bond lengths of hydrogen atoms connected to heavy atoms with an allowable bond length deviation of 10^{-8} Å (**Table 1.2**).

Table 1.2. Parameters used for MD simulations.

Parameters	
Forcefield:	CHARMM 36
Water Model Used:	TIP3P
Timestep:	2fs/step
cut-off (short range electrostatic, Van der wall):	12 Å
Switch distance	11 Å
Rigid bond tolerance between hydrogen and heavy atom	$1*10^{-8}$ Å (ShakeH/LINCS algorithm)
Long Range electrostatic interaction:	Particle Mesh Ewald (PME)
Temperature control:	Langevin dynamics/ velocity rescaling
Temperature:	310 K
Pressure Control:	Langevin Piston/ Parrinello–Rahman barostat
Pressure:	1 atm

Trajectories were saved at 10 ps intervals during the production phase. Multiple independent simulations were considered with different initial velocity distributions, to ensure adequate sampling and convergence. The converged independent trajectories were used for (a) structural analysis, and (b) the final structures were subjected to alchemical free energy calculations, as explained in **section 1.8.3**.

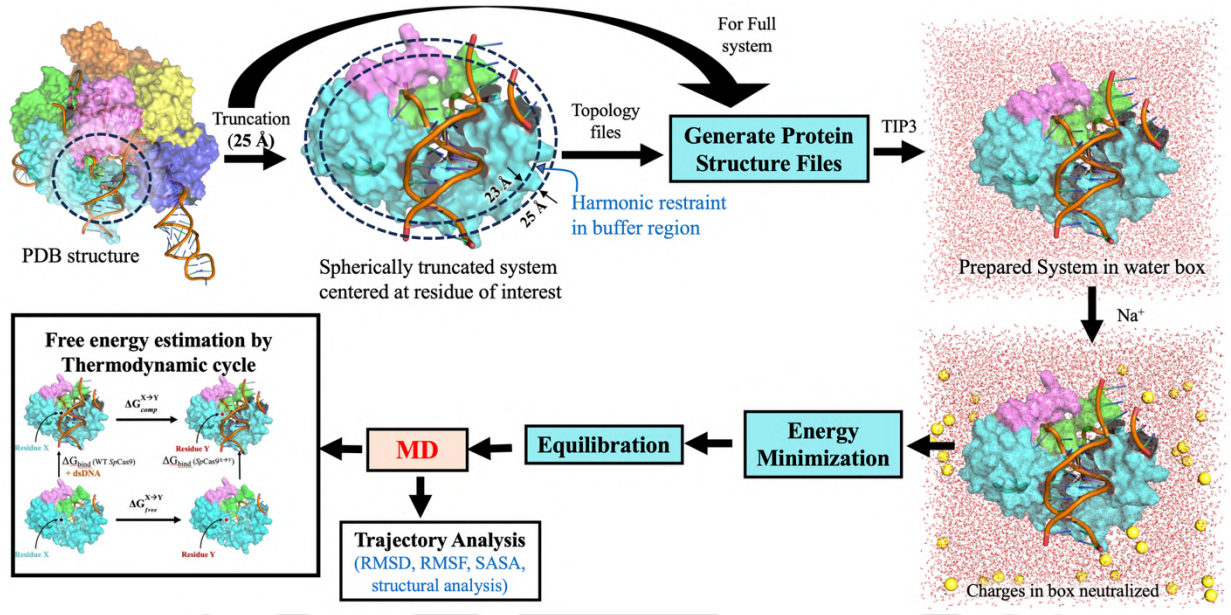


Figure 1.14. Schematic representation of methodology adopted for MD simulations for truncated and full systems. A 25 Å truncated sphere centered at the residue of interest of the precatalytic *Sp*Cas9 was extracted, and the heavy atoms in the outer "buffer region" (23-25 Å) were harmonically restrained to their experimentally determined positions. The system was solvated with a pre-equilibrated water box and charge neutralized by adding monovalent ions. The resulting simulation box is subjected to energy minimization, equilibration (employing NVT followed by NPT ensemble), and production dynamics. The structures obtained from the production molecular dynamics (MD) simulations were used for the relative binding free energy calculations (employing the appropriate thermodynamic cycle).

The MD trajectories are routinely analyzed to extract meaningful insights into structural and dynamic properties of biomolecular systems. Quantities that are routinely analyzed from an MD trajectory include:

(a) **Average structural properties:** Distances, angles, and water distribution are frequently averaged over trajectory frames to characterize specific interactions such as salt bridges or hydrogen bonds, or to assess the local hydration environment. For example, an average distance $\langle d_{ij} \rangle$ between two atoms i and j over N frames are shown in **Equation 1.10**.

$$\langle d_{ij} \rangle = \frac{1}{N} \sum_{t=1}^{N} d_{ij}(t) \tag{1.10}$$

where $d_{ij}(t)$ is the instantaneous distance at frame t.

(b) **Root Mean Square Deviation (RMSD)**: It measures the average deviation of atomic positions from a reference structure (**Equation 1.11**) thus highlights similarity between two structures. It is also often used to monitor structural convergence (plateau of RMSD versus time plot, when no significant structural difference occurs between two time points).

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|\mathbf{r}_i(t) - \mathbf{r}_i^{\mathrm{ref}}\right|^2} \tag{1.11}$$

where $\mathbf{r}_i(t)$ is the position of atom i at time t, $\mathbf{r}_i^{\mathrm{ref}}$ is the reference position, and N is the number of atoms considered.

(c) **Root Mean Square Fluctuations (RMSF)**: It assesses the flexibility of individual residues or atoms by estimating how much a molecule fluctuates from their mean trajectory (**Equation 1.12**).

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T} |\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle|^2} \tag{1.12}$$

where $\langle \mathbf{r}_i \rangle = \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}_i(t)$ is the average position over T frames.

(d) **Radius of Gyration (Rg)**: It provides insights into the compactness or expansion of the molecular structure, useful in folding or unfolding studies.it is defined as the root mean square distance of the atoms/residues from its centre of mass of the biomolecule (**Equation 1.13**).

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i |\mathbf{r}_i - \mathbf{r}_{\mathrm{cm}}|^2}{\sum_{i=1}^{N} m_i}} \tag{1.13}$$

where m_i and $\mathbf{r}_i$ denote the mass and position of atom i, and $\mathbf{r}_{\mathrm{cm}}$ is the center of mass of the molecule.

(e) **Solvent Accessible Surface Area (SASA)**: It quantify exposure of residues to solvent by calculating the surface area of a residue accessible by the solvent.

For truncated system, these properties were estimated only for the heavy atoms of the unrestrained biomolecule except for the buffer region in the boundaries, where restraints were provided.

1.8.3. Thermodynamic Cycle and Relative Binding Energy Estimations

A thermodynamic cycle is a closed loop of linked steps connecting different molecular states (e.g., bound vs. unbound, mutated vs. wild-type), where the net free energy change around the cycle is zero (Hansen & Van Gunsteren, 2014; Shobana et al., 2000). Each step in the cycle represents a reversible transformation between molecular states, which can either be physical changes like substrate binding, or alchemical changes where parts of the molecule are gradually modified or mutated in the simulation. Following this principle, DNA binding to wildtype and mutant *Sp*Cas9 could be described by employing an appropriate thermodynamic cycle (**Figure 1.15**). In this thermodynamic cycle, the vertical arms correspond to physically realizable dsDNA binding to *Sp*Cas9, whereas the horizontal arms denote the alchemical transformation of a wildtype residue in *Sp*Cas9 (say **x**) to a mutant residue (say **y**) either in the complex with dsDNA (upper arm) or in the free *Sp*Cas9 in water (lower arm, **Figure 1.15**). These horizontal arms are unphysical paths and cannot be experimentally realizable (Pohorille et al., 2010). However, free energy changes along these unphysical arms can be estimated computationally with considerable accuracy (Steinbrecher & Labahn, 2010).

The computation of absolute binding free energies along the vertical arms of a thermodynamic cycle is computationally more challenging due to convergence issues (Jorgensen et al., 1988; Lapelosa et al., 2012; N. Singh & Warshel, 2010). Therefore, in this thesis, the relative free energy changes associated with the horizontal (alchemical) arms ($\Delta G_{comp}^{x\to y} - \Delta G_{free}^{x\to y}$) were computed (**Figure 1.15**) and a link between the estimated energetics biomolecule structures were established.

Relative binding free energy is determined by exploiting the path-independence of free energy as a state function, and therefore the net free energy changes across the thermodynamic cycle (**Figure 1.15**) will be zero (**Equation 1.14, 1.15**).

Since free energy is a state function,

$$\Delta G_{comp}^{x\to y} - \Delta G_{\text{bind}}(\text{y}) - \Delta G_{free}^{x\to y} + \Delta G_{\text{bind}}(\text{x}) = 0 \quad (1.14)$$

$$\Delta\Delta G = \Delta G_{comp}^{x\to y} - \Delta G_{free}^{x\to y} = \Delta G_{\text{bind}}(\text{y}) - \Delta G_{\text{bind}}(\text{x}) \quad (1.15)$$

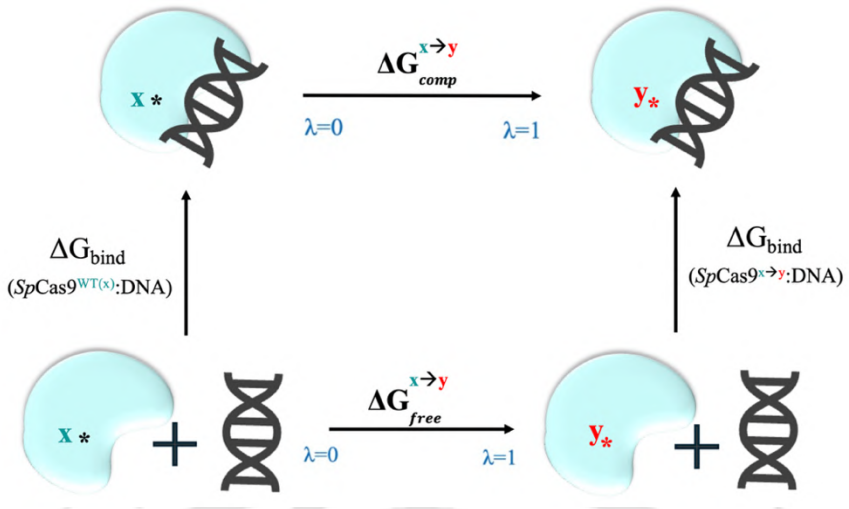


Figure 1.15. Thermodynamic cycle to study the effect of *Sp*Cas9 mutation on DNA binding affinity. Vertical arms represent binding, while horizontal arms depict the alchemical transformation of residue x → y in the *Sp*Cas9, either in complex with dsDNA (above) or free in water (below).

In this thesis, the horizontal arms were calculated by defining a hybrid energy function (U, **Equation 1.16**) along with alchemical coordinate "λ" (McCammon, 1991), where λ=0 and λ=1 represent the endpoints wildtype (x) and mutant (y) respectively.

$$U = \lambda U_y + (1- \lambda) U_x \qquad (1.16)$$

A gradual change in λ from 0 →1 slowly modifies the force field parameters (i.e., electrostatic, van der Waals, and bonded energy term) and transforms the residue **x→y** in the *Sp*Cas9.

Two statistical mechanics approaches were used to calculate the relative binding free energies in this thesis as described below.

1.8.3.1. Free energy perturbation (FEP)

Free energy perturbation (FEP) (Zwanzig, 2004) estimates the free energy change ($\Delta G_{\lambda i \to \lambda i-1}$) between two neighboring "λ" windows (λ_{i-1} and λ_i). The total free energy change along the alchemical path (horizontal arm, i.e., $\Delta G_{comp}^{x \to y}, \Delta G_{free}^{x \to y}$) was estimated by summing over the intermediate (N) states (**Equation 1.17**).

$$\Delta G = \Delta G\,(0 \to 1) = \sum_{i=1}^{N} \Delta G_{\lambda i \to \lambda i-1} = -\beta^{-1} \sum_{i=1}^{N} \ln \langle \exp[-\beta(U_{\lambda i} - U_{\lambda i-1})] \rangle_{\lambda_{i-1}} \qquad (1.17)$$

Where N is the total number of windows. β stands for $1/k_BT$, and k_B and T represent the Boltzmann constant and the temperature, respectively.

1.8.3.2. Bennett Acceptance Ratio (BAR)

The free energy change ($\Delta G_{\lambda i \rightarrow \lambda i-1}$) between two neighboring "λ" windows (λ_{i-1} and λ_i) can be obtained by calculating Bennett Acceptance Ratio (BAR) (Bennett, 1976; **Equation 1.18**).

$$\Delta G_i = -\beta^{-1} \ln \left(\frac{\langle 1+e^{-\beta(\Delta U \Delta \lambda_i - C_i)} \rangle_{i+1}}{\langle 1+e^{+\beta(\Delta U \Delta \lambda_i - C_i)} \rangle_i} \right) + C_i \quad (1.18)$$

Where, $\Delta U \Delta \lambda i$ is the potential energy difference between adjacent window (i and i+1), the angled brackets $\langle \cdots \rangle_i$ and $\langle \cdots \rangle_{i+1}$ denote ensemble averaging at windows i and i+1, respectively, and Ci is an adjustable constant, iteratively optimized to make the two ensemble averages equal. The BAR method uses simulation data from both windows (forward and backward direction) to provide a statistically optimal estimate of the free energy difference (Bennett, 1976). For each window, the energy difference is computed using samples from both ensembles, and the constant Ci is refined until the averages converge. This yields a robust and accurate free energy estimates.

In this thesis, these calculations estimated ΔG by using a minimum of 51 and a maximum of 201 equally spaced λ points along the horizontal arm from 0 to 1 in increments of 0.02-0.005, depending on the size of alchemical transformation. Individual λ windows were simulated for a minimum of 3 ns, a maximum of 10 ns, out of which the initial 1 ns of the trajectory at every lambda window were treated as equilibration (time required to adjust to the new Hamiltonian) and were excluded from free energy estimations (Pohorille et al., 2010). To ensure smooth transitions and avoid numerical instabilities at the end points of the alchemical λ simulations, a soft-core potential was employed (Beutler et al., 1994). Free energy was estimated are usually estimated for forward (λ= 0 →1) and backward (λ= 1→0) transformation. The forward and backward free energies associated with alchemical transformations were averaged and reported as $\Delta G_{comp}^{x \rightarrow y}$ and $\Delta G_{free}^{x \rightarrow y}$, along with the associated statistical error. The relative binding free energy ΔΔG was calculated as $\Delta\Delta G = \Delta G_{comp}^{x \rightarrow y} - \Delta G_{free}^{x \rightarrow y}$ and the errors for ΔΔG were obtained by propagating the errors associated with $\Delta G_{comp}^{x \rightarrow y}$ and $\Delta G_{comp}^{x \rightarrow y}$. The sign of ΔΔG (negative/positive) suggests that the DNA binding is favoured/disfavoured in response to *Sp*Cas9 mutation. The

magnitude of the calculated ΔΔG is attributed to the strength of the preference. To ensure robust sampling, alchemical free energy calculations were performed with multiple independent simulation replicas, typically 3-5 runs per transformation. Convergence was assessed by (i) monitoring overlap between forward and reverse free energy distributions, (ii) plateau in the ΔG versus time plot, (iii) consistency of ΔG estimates across ≥3 independent replicas, (iv) matching ΔΔG values across truncated (25 Å and 30 Å radius) and full systems.

Alchemical transformation is a standard, well-established procedure for evaluating binding affinities in biomolecular systems, particularly for protein–ligand interactions (Garg & Debnath, 2025; Khalak et al., 2021; Muegge & Hu, 2023). In this work, we have extended the framework to protein and nucleic acid systems. While the CHARMM36 force field provides a comprehensive description of biomolecular interactions, classical force fields inherently lack explicit electronic polarization, which can bias conformational sampling (Jing et al., 2019; Riniker, 2018; Vanommeslaeghe & Mackerell, 2014). However, relative binding affinity estimation through alchemical free energy calculations (as adopted in this thesis) is often more accurate (typically within 1 kcal/mol error) than absolute binding affinity predictions (Bhati et al., 2017; Mey et al., 2020). Note that both absolute and relative affinity calculations depend on the quality of the force field, biomolecular flexibility, and system preparation (including solvation, ions, etc.). But relative affinity methods are less sensitive to such issues, especially when transforming ligands (DNA bases in our case) that are chemically similar and bind in comparable poses (Molani & Cho, 2024).

1.8.4. Software used in this thesis

In this thesis, two open-source MD tools: NAMD (Phillips et al., 2005) and GROMACS (Spoel et al., 2005) were used to perform MD simulations and analyze the results. The hybrid DNA structures and topologies were generated with the help of the PMX package (Gapsys et al., 2015) or mutator plugin of VMD. Structural visualization and trajectory analysis were carried out using PyMOL (DeLano et al., 2002) and VMD (Humphrey et al., 1996). Trajectory data were plotted using Microsoft Excel, the Grace plotting tool, and custom scripts written in Python and R. VMD, Tcl and Unix shell scripting were routinely used for post processing MD trajectories. All simulations were executed on the high-performance computing clusters: Param Ishan and Param Kamrupa at the Supercomputing Facility, IIT Guwahati.

Chapter 2

Effect of Single E1219V Mutation on the Energetics of PAM Recognition

This chapter is published in J. Chem. Inf. Model, 2024, 64, 8, 3237–3247

Popular RNA-guided DNA endonuclease, Cas9 from *Streptococcus pyogenes* (*Sp*Cas9) recognizes the canonical 5'-NGG-3' protospacer adjacent motif (PAM) and triggers double-stranded DNA cleavage activity. Mutations in *Sp*Cas9 demonstrated to expand the PAM readability and hold promise for therapeutic and genome editing applications. However, the energetics of the PAM recognition and its relation to the atomic structure remains unknown. Using X-ray structure (pre-catalytic *Sp*Cas9:sgRNA:dsDNA) as a template, we calculated the change in the PAM binding affinity in response to *Sp*Cas9 mutations using computer simulations. E1219V mutation in *Sp*Cas9 fine-tunes the water accessibility in the PAM binding pocket and promotes new interaction between *Sp*Cas9:non-canonical T-rich PAM, thus, weakening the PAM stringency. Nucleotide-specific interaction of two arginine residues (i.e., R1333 and R1335 of *Sp*Cas9) ensured stringent 5'-NGG-3' PAM recognition. R1335A substitution (*Sp*Cas9^{R1335A}) completely disrupts the direct interaction between *Sp*Cas9 and PAM sequences (canonical or non-canonical), accounting for the loss of editing activity. Interestingly, double-mutant (*Sp*Cas9^{R1335A,E1219V}) boosts DNA binding affinity by favouring protein:PAM electrostatic contact in a desolvated pocket. The underlying thermodynamics explains the varied DNA cleavage activity of *Sp*Cas9 variants. A direct link between the energetics, structures, and activity is highlighted, which can aid the rational design of improved *Sp*Cas9-based genome editing tools.

2.1. Background

CRISPR/Cas9 system have revolutionized genome editing, offering unprecedented opportunities for precise and efficient manipulation of DNA sequences (Adli, 2018). The single Cas9 protein in complex with a single guide RNA (sgRNA) can recognize and catalyze site-specific double-stranded DNA (dsDNA) cleavage, making it a simple but powerful tool for gene editing (Jinek et al., 2012). This site-specific Cas9 catalyzed cleavage in the dsDNA requires (a) the presence of Protospacer Adjacent Motif (PAM; 2 to 6 nucleotide sequence) in the non-target strand of the DNA (ntDNA) and (b) 20 nucleotide complementarities between sgRNA and target strand of the DNA (tDNA). The PAM-interacting domain (PID) of Cas9 recognizes the PAM sequence in the non-target strand and triggers the DNA unwinding, followed by sgRNA:tDNA base-pairing to form the precatalytic complex (**Figure 2.1**). Cas9 catalyzes the site-specific dsDNA cleavage (three nucleotides upstream of the PAM) by involving two nuclease domains (HNH and RuvC, **Figure 2.1**) in the presence of Mg^{+2} ions (Jiang and Doudna, 2017a; Rath et al., 2015). The 5'-NGG-3' serves as the canonical/cognate PAM sequence for the Cas9 of *Streptococcus pyogenes* (*Sp*Cas9), which not only prevents self-DNA cleavage in bacteria but also identifies the correct location for the dsDNA cleavage in the foreign DNA (Jiang and Doudna, 2017a; Rath et al., 2015). Key residues of the PAM interacting domain (PID) of *Sp*Cas9 include R1335 (stabilized by E1219 salt bridge interaction) and R1333, which ensure specificity by forming stable base-specific interactions with two guanine bases G2 and G3 of the canonical 5'-NGG-3' sequence (**Figure 2.1**) (Anders et al., 2014).

Despite its remarkable potential, the applicability of the promising *Sp*Cas9 in genome editing is limited by its stringent requirement for a specific PAM sequence (i.e., 5'-NGG-3') adjacent to the target site, restricting its application to a limited range of genes of interest (Guo et al., 2019; Kleinstiver et al., 2015). Hu et al. designed xCas9 3.7 variant (contain seven point mutations: A262T, R324L, S409I, E480K, E543D, M694I and E1219V) with relaxed PAM specificities (Nishimasu et al., 2018). Guo et al. reported that a single E1219V mutation in *Sp*Cas9 expanded the PAM recognition by allowing cleavage activity for various noncanonical PAM sequences, 5'-GAT-3' or 5'-TGT-3' including the wildtype 5'-TGG-3. They hypothesized that the broader PAM

compatibility in the E1219V variant of *Sp*Cas9 was attributed to the unrestricted rotamerization of R1335 residue (Guo et al., 2019). Despite the advancement in the structural/biochemical investigations, the mechanism of PAM recognition in terms of energetics and its link to *Sp*Cas9 mutation was still unknown.

In this chapter, we address the effect of E1219V mutation in the wild-type and R1335A-substituted *Sp*Cas9 (*Sp*Cas9^{R1335A}) on the energetics of dsDNA binding by molecular dynamics free energy calculations. We estimated the change in dsDNA binding affinity upon E1219 → V1219 mutation in the *Sp*Cas9 and *Sp*Cas9^{R1335A} variants. Various dsDNA sequences differing in their PAM sequences (5'-TGG-3' or 5'-TGT-3' or 5'-GAT-3' or 5'-TTG-3' or 5'-TTT-3') were considered. The results showed that E1219V improves the binding affinity to T-rich non-canonical PAM sequences by creating a hydrophobic pocket for the thymine nucleotide and promoting new PAM-PID interactions (base-specific and or base-non-specific) in both wild-type *Sp*Cas9 and *Sp*Cas9^{R1335A} variant. The results not only provided the link between atomic structure and the energetics of PAM selectivity by *Sp*Cas9 but also explained the previous experimental observations (Guo et al., 2019; Nishimasu et al., 2018) and deciphered the mechanisms of PAM specificity.

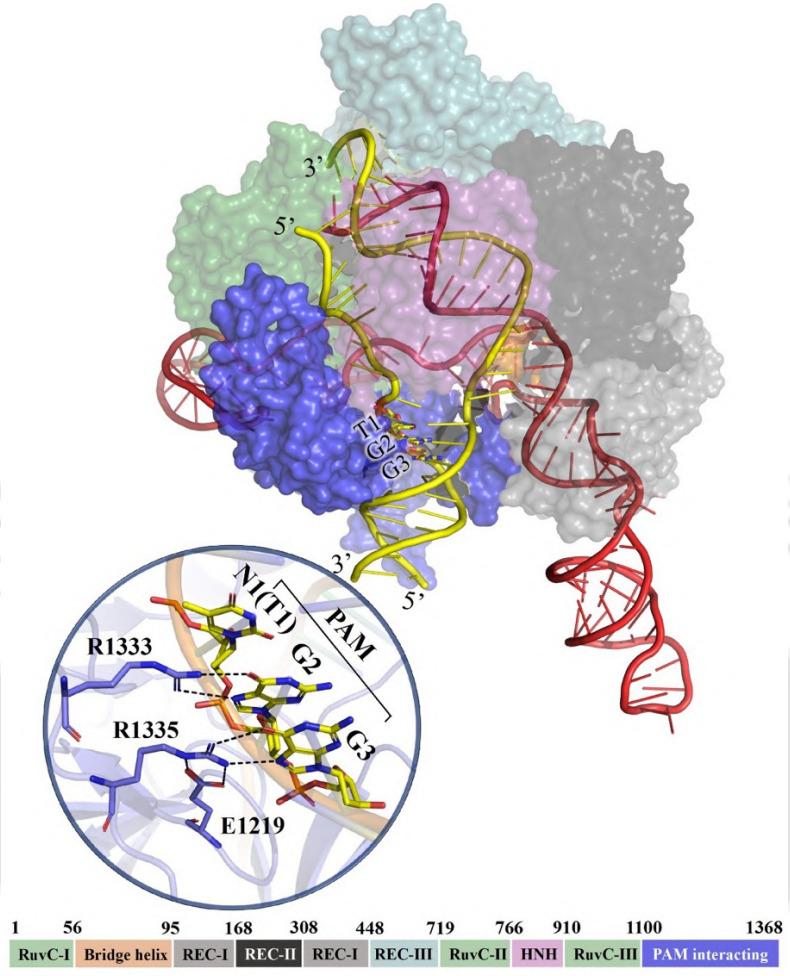


Figure 2.1. Pre-catalytic *Sp*Cas9 (surface) in complex with sgRNA (red) and dsDNA (yellow) in the absence of Mg^{2+} (PDB 5F9R). Distinct domains were coloured differently. Zoomed-in view highlighting the interaction between PAM (5'-TGG-3') and PID. Key residues were shown in sticks, and the interaction network was shown in dotted lines. Hydrogens were kept hidden for clarity. A color-coded schematic diagram of *Sp*Cas9 domain organization is shown at the bottom.

2.2. Methodology

2.2.1. Molecular Dynamics (MD) Setup

The structure of the precatalytic *Sp*Cas9 bound to sgRNA and dsDNA (containing canonical 5'-TGG-3' PAM sequence) was retrieved from Protein Data Bank (PDB 5F9R, resolution = 3.4 Å)

(Jiang et al., 2016). A 25 Å spherically truncated system centered at the E1219 was extracted from the PDB structure, which predominantly encompasses the PAM-interacting domain (PID) along with smaller segments of the RuvC, HNH, and REC domains (**appendix Figure A2.1**). The resulting structure was solvated in an explicit water box, and neutralized with counter (Na^+/Cl^-) ions (**appendix Table A2.1**). The resulting system was subjected to energy minimization, equilibration, and production dynamics, with a protocol elaborated in chapter 1 (**section 1.8.2**). Simulations were conducted at a temperature of 310 K and a constant pressure of 1 bar. Langevin Dynamics (Phillips et al., 2005) and a Langevin Piston (Feller et al., 1995; Martyna et al., 1994) were used to control temperature and pressure, respectively. Models of precatalytic *Sp*Cas9 bound to non-canonical PAM sequences (viz., 5'-TGT-3' or 5'-GAT-3' or 5'-TTG-3' or 5'-TTT-3' PAM) were generated by appropriate substituting in the dsDNA. R1335 was substituted with A1335 in the *Sp*Cas9 to model precatalytic *Sp*Cas9^{R1335A} complexes with various PAM sequences (5'-TGG/TGT/GAT/TTG/TTT-3'). After the production dynamics, the final structure of the complex was subjected to E1219→V1219 alchemical transformation for estimating relative binding free energies (described in the next section). The convergences of the MD structures and the estimated free energies were confirmed by performing the simulations for a bigger 30 Å truncated system and a full *Sp*Cas9 system (**appendix Table A2.1**). The calculations were repeated multiple times for the smaller truncated system (25 Å model) to minimize the computational cost and ensure adequate sampling and, the results were compared with the truncated systems (**appendix Table A2.2-2.10**).

2.2.2. Relative Binding Free Energy Calculations

The final MD structures obtained from conventional MD simulations (**section 2.2.1, appendix Table A2.1**) were subjected to the alchemical transformation, where amino acid residue E1219 is transformed to V1219 residue (in the *Sp*Cas9). The binding free energy differences (ΔΔG) of *Sp*Cas9 binding to dsDNA containing different PAM sequences in response to E1219V mutation were estimated using an appropriate thermodynamic cycle (**Figure 2.2a**), where vertical arms represent dsDNA binding and horizontal arms denote the alchemical transformation of E1219 to V1219 either in the complex with dsDNA (upper arm) or in the free *Sp*Cas9 in water (lower arm,

Figure 2.2a). In this chapter, the free energy changes along the horizontal arms (ΔG_{comp}) and (ΔG_{free}) were computed using the Bennett Acceptance Ratio (BAR) and Free Energy Perturbation (FEP) (Zwanzig, 2004) methods. The detailed principles of these approaches are explained in chapter 1 (section 1.8.3). Alchemical transformations were performed using 51 λ windows spaced from 0 to 1 in 0.02 increments. Each window was simulated for 3–10 ns depending on system size, with the first 1 ns discarded as equilibration. The forward (λ = 0→1) and backward (λ = 1→0) free energy differences associated with the alchemical transformations (ΔG_{comp} and ΔG_{free}), along with the associated statistical error were estimated by the ParseFEP plugin of VMD (Humphrey et al., 1996; Liu et al., 2012) by employing both the Free energy Perturbation (FEP) (Zwanzig, 2004) and BAR estimations (Bennett, 1976; Liu et al., 2012).

2.2.3. Sampling and Convergence

A popular choice for structural comparison and assessment of convergence is computing the root-mean-square deviation (RMSD). RMSD plot of the protein's heavy atoms (**appendix Figure A2.2**) showed that the MD structure reached a state of convergence after approximately ~30 ns. Convergence was achieved when the RMSD values were stabilized (the RMSD plot reached a plateau), indicating minimal fluctuations (< 0.1 Å) in the protein's structure. Post convergence, the average RMSD was 1-1.5 Å from the template X-ray structure, indicating a close resemblance of MD structures with X-ray structures. The alignment of the MD structure after 100 ns with the X-ray structures (**appendix Figure A2.3**) demonstrated high structural similarity, and zooming into the PAM binding region, the interaction pattern between *Sp*Cas9 and 5'-TGG-3' PAM sequence was fully preserved after MD. All the structures preserved a total of four base-specific hydrogen bonding between the 'GG' residues of 5'-TGG'3' PAM and *Sp*Cas9 residues of the PAM Interacting (PID) domain (R1333 and R1335). Even after extending the simulation to 500 ns, the plateau in the RMSD plots (with a small fluctuation < 0.1 Å) were observed (**appendix Figure A2.4**), which indicated that simulations more or less reproduced the experimentally characterized PAM binding pocket. Thus, we may argue that the simulations are sampling the relevant conformations of the protein around the minima (or X-ray structure) in the potential energy

hypersurface. Moreover, the PAM binding pocket is found to be more or less identical for different simulation models (full versus truncated models, (**appendix Table A2.3**). Consequently, it substantiates the reliability and fidelity of our MD simulations in sampling the desired minima of the potential energy hypersurface.

MD Structures of *Sp*Cas9 bound to 5'-TGT/GAT/TTG/TTT-3' PAM revealed fewer *Sp*Cas9:PAM interactions. Compared to four hydrogen bonds between *Sp*Cas9 residues (R1333 and R1335) and 5'-TGG-3' PAM, only two hydrogen bonds were observed between R1333 and 2nd 'G' residue of 5'-TGT-3' PAM, while no interactions were present between the R1335 residue and PAM sequence (**Figure 2.3c**). Similarly, two base-specific hydrogen bonds were observed between R1335 and 3rd 'G' of 5'-TTG-3' PAM, with no interactions with R1333 residues (**Figure 2.4a**). Moreover, no *Sp*Cas9:PAM interactions were observed in the case of 5'-GAT-3' and 5'-TTT-3' PAM (**Figure 2.3e, 2.4c**), suggesting that 5'-GAT-3' and 5'-TTT-3' PAM as a PAM sequence is completely unrecognizable by wild-type *Sp*Cas9. In all cases, the E1219 residue forms stable interactions with the side chain of the R1335 residue. These observations are supported by the root mean square fluctuation (RMSF) plot (**appendix Figure A2.5a-c**), which highlighted an increase in flexibility around PAM interacting residues (residue R1333 and R1335) in the case of *Sp*Cas9 bound to non-canonical PAM, where the interactions are broken. R1335A mutations were observed to break all the *Sp*Cas9:PAM interactions and made the R1333 residue notably more flexible (**appendix Figure A2.5b**). A total of ~8.5 μs of MD trajectory were used for structural analysis and ensuring convergence.

Multiple independent alchemical simulations (up to five replicas, (**appendix Table A2.2**) were performed with varying sampling strategies (5 ns and 10 ns per lambda window). The statistical error was computed from the degree of overlap of the underlying probability distributions between the two neighboring windows (Chipot & Pohorille, 2007; Liu et al., 2012; Shell et al., 2007) (**appendix Figure A2.6**). Sufficient overlap between adjacent states and low statistical error values validated the choice of lambda windows and the reliability of ΔΔG values. A total of ~19 μs of alchemical free energy simulations have been performed to achieve good convergence and reasonable statistical error. The estimated free energies from different independent trials are in

excellent agreement (ΔG differs by < 0.5 kcal/mol, **appendix Table A2.2**). The estimated ΔG's from various independent trials were averaged and reported as ΔG_{comp} and ΔG_{free}, and the standard error of the mean (s.e.m) was reported as the errors. The errors associated with the ΔG_{comp} and ΔG_{free} were propagated and reported as errors associated with the relative free energy (ΔΔG) (**Figure 2.2b**). Estimated energetics were shown to be practically insensitive to the size of the simulation system (**appendix Table A2.2**). The negative/positive sign of ΔΔG indicates PAM containing dsDNA binding to *Sp*Cas9 is favoured/disfavoured by E1219V mutation, respectively, while the magnitude of ΔΔG denotes the strength of the preference.

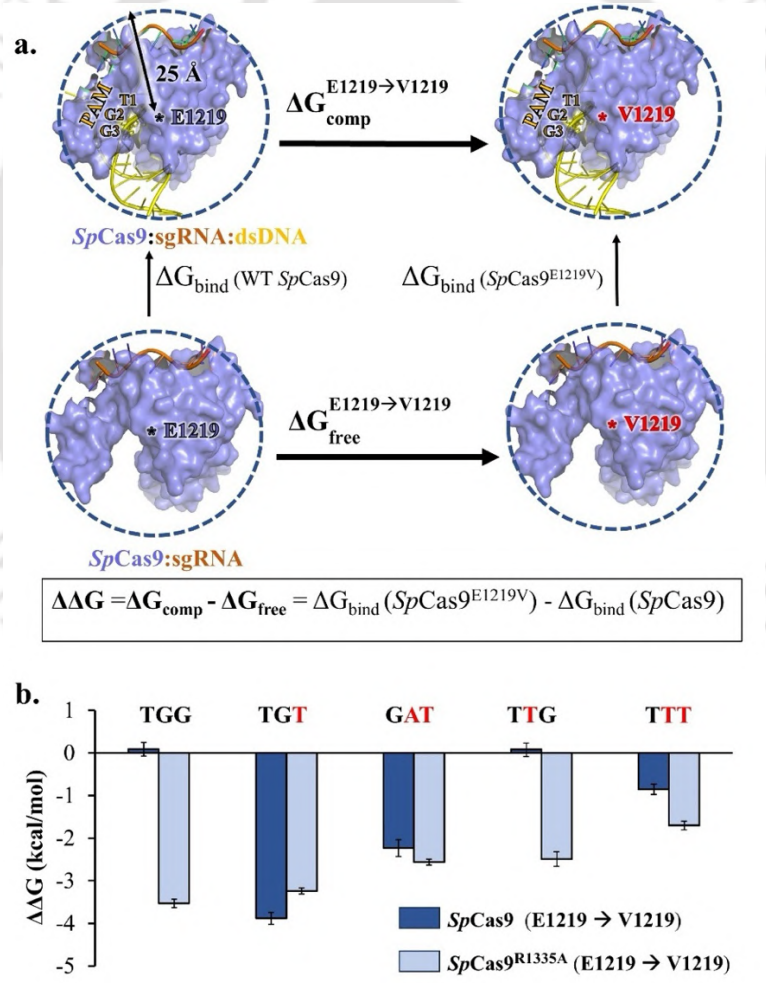


Figure 2.2. (**a**) Thermodynamic cycle for estimating the effect of E1219/V1219 mutation on the *Sp*Cas9 binding to dsDNA. The relative free energy (ΔΔG) is calculated as $\Delta\Delta G = \Delta G_{comp} - \Delta G_{free} = \Delta G_{bind}$

(*Sp*Cas9^{E1219V}) – ΔG_{bind} (WT *Sp*Cas9) (**b**) Calculated changes in binding free energy, for 5'-TGG-3', 5'-TGT-3', 5'-GAT-3', 5'-TTG-3' and 5'-TTT-3' in response to E1219/V1219 mutation in the *Sp*Cas9 (light-blue) and *Sp*Cas9^{R1335A} (dark-blue). Error bars, 1 s.e.m.

To validate the reliability of ΔΔG computed by spherically truncated models, we have performed computationally intensive alchemical simulation by considering the full system in a box of water (Total sampling ~ 1 μs) without applying harmonic restraint. Two different systems (**appendix Table A2.1-A2.4)** differing in their PAM sequence have been considered *Sp*Cas9:sgRNA:DNA^{TGG} (Size: 210509 atoms) and *Sp*Cas9:sgRNA:DNA^{TGT} (Size: 211200 atoms). The values of ΔΔG as well as trajectory averaged key distances in PAM interacting pocket were compared with the results obtained from the truncated models (25Å and 30Å). Clearly, the estimated energetics (ΔΔG) and the structural integrity of the PAM binding pocket are independent of the size of the system (similar values of ΔΔG in both truncated and full systems; **appendix Table A2.2**). The robustness of the calculated relative binding affinity (ΔΔG) and the structural integrity of the PAM binding pocket is evident (**appendix Table A2.2-A2.4**).

2.3. Results

2.3.1. Effect of E1219→V1219 mutation on the energetics of *Sp*Cas9 binding to various PAM sequences

To decipher the energetic origin of PAM recognition by *Sp*Cas9, we carried out classical molecular dynamics free-energy calculations on wild-type (*Sp*Cas9) and R1335A-substituted *Sp*Cas9 (SpCas9^{R1335A}) in complex with various DNA substrates differing in their PAM sequence (5'-TGG or TGT or GAT-3' or 5'-TTG-3' or 5'-TTT-3'). The pre-catalytic state of wild-type *Sp*Cas9 in complex with sgRNA and canonical 5'-TGG-3' PAM sequence (PDB 5F9R) (Jiang et al., 2016) was used as a template for computational analysis (**Figure 2.1**). The two arginine residues i.e., R1333 and R1335 of the PAM interacting domain (PID) of the wild-type (WT) *Sp*Cas9, form hydrogen bonds with the Hoogsteen edge of the 2nd and 3rd guanine of 5'-TGG-3' PAM sequence, which ensures specificity (**Figure 2.1**). Salt-bridge interaction between E1219 and R1335

indicates charge neutralization and employs restraint to R1335 residue in the PAM binding pocket (**Figure 2.1**). Models of the wild-type *Sp*Cas9 in complex with 5'-TGT-3', 5'-GAT-3', 5'-TTG-3', and 5'-TTT-3' were generated by modifying the canonical double-stranded 5'-TGG-3' sequences. Relative binding free energies between wild-type *Sp*Cas9 and E1219V-variant of *Sp*Cas9 (*Sp*Cas9^{E1219V}) for different PAM sequences (5'-TGG-3' or 5'-TGT-3' or 5'-TAG-3' or 5'-TTG-3' or 5'-TTT-3') were estimated by employing E1219 → V1219 alchemical transformation methodology described in **Figure 2.2a**. The calculations estimated the change in binding affinity between *Sp*Cas9 of dsDNA substrate (containing 5'-TGG/TGT/GAT/TTG/TTT-3' PAM sequences) upon E1219V mutation.

The results are summarized in **Figure 2.2b** and **appendix Table A2.2**. *Sp*Cas9 and *Sp*Cas9^{E1219V} bind to canonical 5'-TGG-3' PAM sequence with more or less similar affinities ($\Delta\Delta G \sim 0$ kcal/mol, **Figure 2.2b**), corroborating the previous experimental observation (similar catalytic activities of *Sp*Cas9 and *Sp*Cas9^{E1219V} for TGG containing DNA substrate)(Guo et al., 2019). Hydrogen bonding between the arginine (R1333 and R1335) and G-rich canonical 5'-TGG-3' PAM was found to be preserved in response to E1219V mutation in the *Sp*Cas9 (**Figure 2.3a,b**, and **appendix Table A2.3**). The E1219V mutation disrupted the electrostatic interaction between E1219 and R1335 and certainly expanded the rotameric conformational space for the side-chain of R1335 (primarily in the substrate-free *Sp*Cas9, supported by an increase in RMSF in **appendix Figure A2.5c**), enabling it to adopt alternate conformations, but the same does not play any role in 5'-TGG-3' binding affinity ($\Delta\Delta G \sim 0$ kcal/mol).

On the other hand, *Sp*Cas9^{E1219V} prefers binding to the non-canonical 5'-TGT-3' and 5'-GAT-3' sequences (the former being noticeable) relative to the wild-type *Sp*Cas9 (negative value to $\Delta\Delta G$, **Figure 2.2b**). Experiments indeed confirmed significant improvement in DNA cleavage activity for substrates containing non-canonical 5'-TGT-3' and 5'-GAT-3' sequences (the former being prominent) in response to E1219V mutation in the *Sp*Cas9 (Guo et al., 2019). The hydrogen bonding between R1335 and the PAM sequence was absent in the *Sp*Cas9:dsDNA(TGT) complex (**Figure 2.3c**, and **appendix Table A2.4**). However, the hydrogen bonding interaction between R1333 and 2nd guanine was preserved in both the *Sp*Cas9 and *Sp*Cas9^{E1219V} (**Figure 2.3c,d**, and

appendix Table A2.4). The stabilizing effect of E1219V mutation in *Sp*Cas9:dsDNA(TGT) is threefold (**Figure 2.3d**). First, the C5-methyl of the 3rd thymine nucleotide is enclosed in a hydrophobic environment created by the aliphatic side chain of V1219. Second, it establishes hydrogen bonding between R1333 and 3rd thymine, and third, perhaps the most important from the energetic viewpoint, promote a non-specific salt-bridge interaction between R1335 and DNA backbone. The difference in binding free energies between *Sp*Cas9 and *Sp*Cas9^{E1219V} bound to the non-canonical 5'-TGT-3' PAM was about 4 kcal mol^{-1}, which corresponds to a ~1000-fold increase in affinity for the latter.

The absence of hydrogen bonding between the 5'-GAT-3' and the PID domain of WT *Sp*Cas9 indicates weak binding (**Figure 2.3e**, and **appendix Table A2.5**). Distinctly different rotameric conformation of R1335 was observed in *Sp*Cas9:dsDNA(GAT) complex relative to *Sp*Cas9:dsDNA (TGG/TGT) complexes (**Figure 2.3a,c,e**). The positively charged R1335 is repelled by the -NH_2 group present in the 2nd Adenine (A2) for obvious electrostatic reasons. E1219V mutation in the *Sp*Cas9:dsDNA(GAT) complex improves the binding affinity by ~2 kcal/mol (**Figure 2.2b**) by promoting R1333:thymine interaction (**appendix Table A2.5**) and solvent exclusion around the methyl group (C5 atom) of 3rd thymine (**Figure 2.3f**). Direct interaction between the side-chain of R1335 and 5'-GAT-3' was absent in both the *Sp*Cas9:dsDNA(GAT) and *Sp*Cas9^{E1219V}:dsDNA(GAT) complexes (**Figure 2.3e,f**).

Single E1219V mutation displayed a significantly enhanced cleavage activity as compared to the wild-type *Sp*Cas9 in substrates containing 5'-TGT-3' and 5'-GAT-3' PAM, former being relatively prominent (Guo et al., 2019), in line with the estimated energetics ($\Delta\Delta G$ ~ -4 kcal/mol and -2 kcal/mol for substrates containing 5'-TGT-3' and 5'-GAT-3' PAM respectively, **Figure 2.2b**). The estimated free-energy differences (**Figure 2.2b**) corroborate very well with the catalytic activity of WT and related variants of *Sp*Cas9 (Guo et al., 2019), thus providing the thermodynamic origin of PAM specificity.

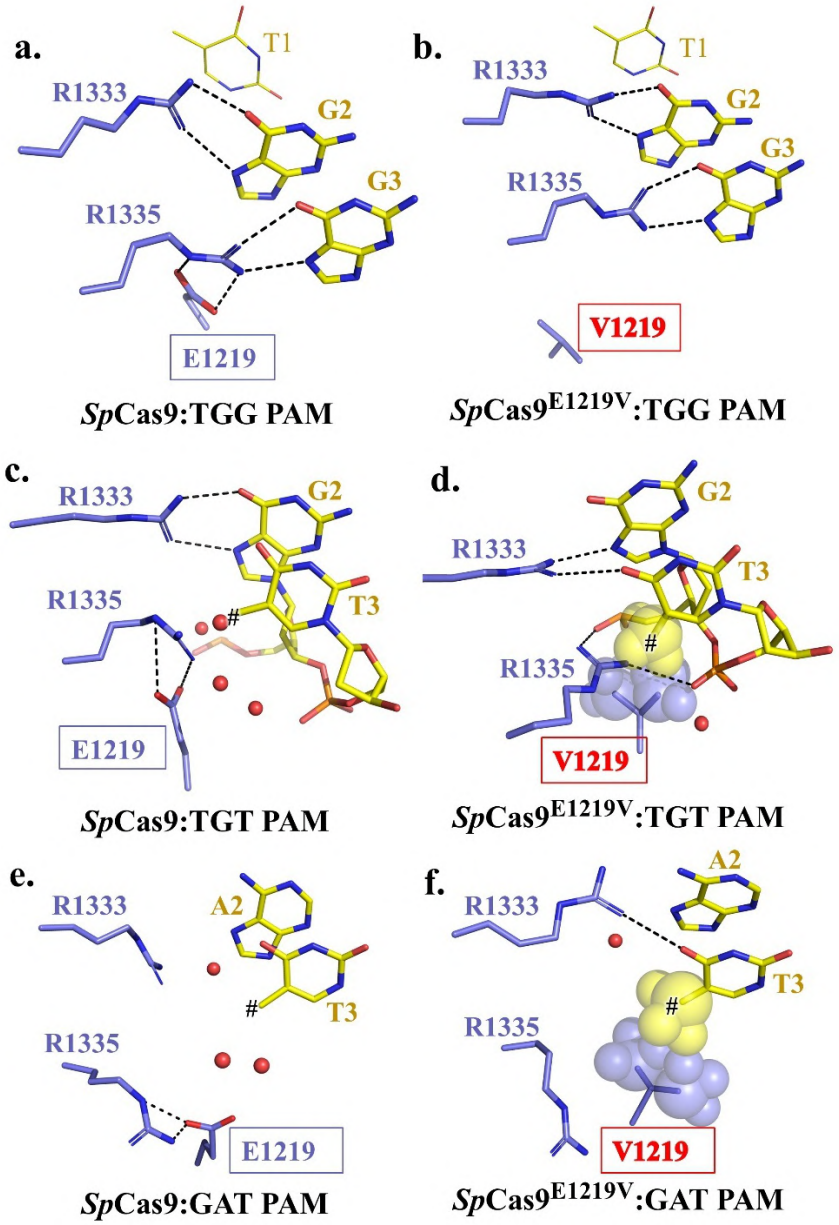


Figure 2.3. Structural comparison of PAM binding pocket in the wild-type *Sp*Cas9 (left) and *Sp*Cas9^{E1219V} (right) pre-catalytic complex. (**a, b**) 5'-TGG-3' binding pocket, **(c, d)** 5'-TGT-3' binding pocket, and **(e, f)** 5'-GAT-3' binding pocket. Key residues are represented in sticks, and the interaction network is highlighted in dotted lines. Hydrogen atoms are not shown for clarity. Hydrophobic pocket created by V1219 and the aliphatic CH_3 group of 3rd thymine is highlighted with transparent sphere representation. Water molecules present within 4 Å of the CH_3 group (# is the C5 atom) of the 3rd thymine are shown in red spheres.

The relative binding free energy calculations were also extended to *Sp*Cas9 bound to other T-rich non-cognate PAM sequences like 5'-TTG-3' and 5'-TTT-3'. E1219→V1219 mutation in *Sp*Cas9 does not alter the binding affinity to 5'-TTG-3' PAM (ΔΔG ~ 0 Kcal/mol, **Figure 2.2b**) since the protein: PAM interaction was preserved (R1335 and the third guanine of 5'-TTG-3', **Figure 2.4a,b** and **appendix Table A2.6**). No direct interaction between the *Sp*Cas9 and 5'-TTT-3' sequence was observed (**Figure 2.4c,d**). However, E1219 →V1219 mutation weakly prefers 5'-TTT-3' PAM binding (**Figure 2.4b)** and promotes *Sp*Cas9^{E1219V}:PAM interactions (**Figure 2.4a,b** and **appendix Table A2.7**).

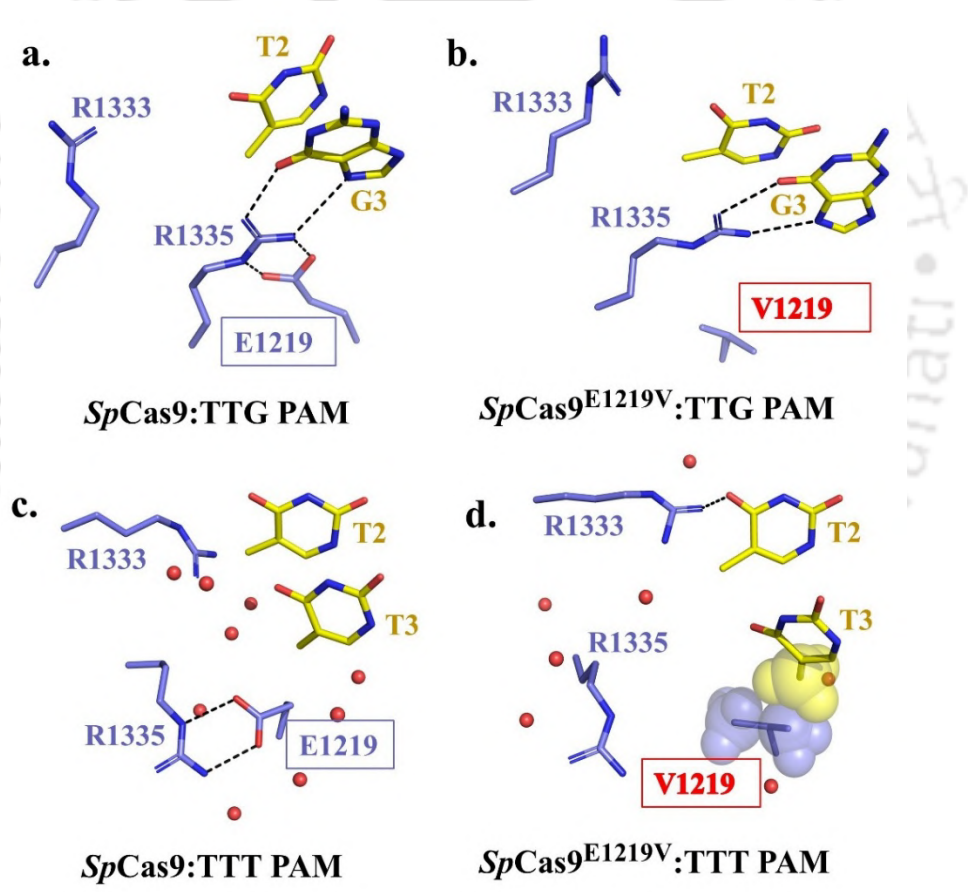


Figure 2.4. Structural comparison of PAM binding pocket in the wild-type *Sp*Cas9 (left) and *Sp*Cas9^{E1219V} (right) pre-catalytic complex. **(a, b)** 5'-TTG-3' binding pocket, **(c, d)** 5'-TTT-3' binding pocket. Hydrophobic pocket created by V1219 and the aliphatic CH_3 group of 3rd thymine is highlighted with transparent sphere representation.

2.3.2. Effect of E1219→V1219 mutation on the energetics of *Sp*Cas9^{R1335A} binding to various PAM sequences

The effect of critical R1335 in PAM selectivity was examined by repeating the calculations where R1335 is substituted to an A1335 amino acid in the *Sp*Cas9. R1335A substitution in the *Sp*Cas9 destabilizes the complex by disrupting the hydrogen bonding between R1333 and PAM sequences (5'-TGG/TGT-3') (**Figure 2.5a**). In vitro cleavage activity assay of purified wild-type *Sp*Cas9 and the *Sp*Cas9^{R1335A} for a target plasmid (containing TGG PAM sequence) confirmed efficient cleavage by wild-type *Sp*Cas9, whereas *Sp*Cas9^{R1335A} showed almost no activity (Nishimasu et al., 2018) in line with MD structures. Modification at the 1335 position was found to have a surprisingly large effect on the E1219/V1219 *Sp*Cas9 selectivity on 5'-TGG-3' bound *Sp*Cas9. E1219→V1219 mutation is strongly favoured for the canonical 5'-TGG-3' binding in the *Sp*Cas9^{R1335A} variant ($\Delta\Delta G \sim -3.5$ Kcal/mol, **Figure 2.2b**), which is otherwise unselective for the unmodified variant (*Sp*Cas9). E1219V mutation in *Sp*Cas9^{R1335A} mutants also favoured other non-canonical (5'-TGT/GAT/TTT-3') PAM binding by similar magnitudes as of unsubstituted *Sp*Cas9 (**Figure 2.2b**). These preferences were attributed to the desolvation of the PAM binding pocket by the E1219V mutation (**appendix Table A2.9, A2.10**), which promotes the formation of base-specific hydrogen bonding between R1333 and the PAM nucleotide(s) (**Figure 2.5b, appendix Figure A2.7** and **appendix Table A2.8-A2.10**). A similar effect on 5'-TTG-3' PAM selectivity was also evident for E1219→V1219 mutation in *Sp*Cas9^{R1335A} (**Figure 2.2b, appendix Figure A2.7** and **appendix Table A2.9, A2.10**).

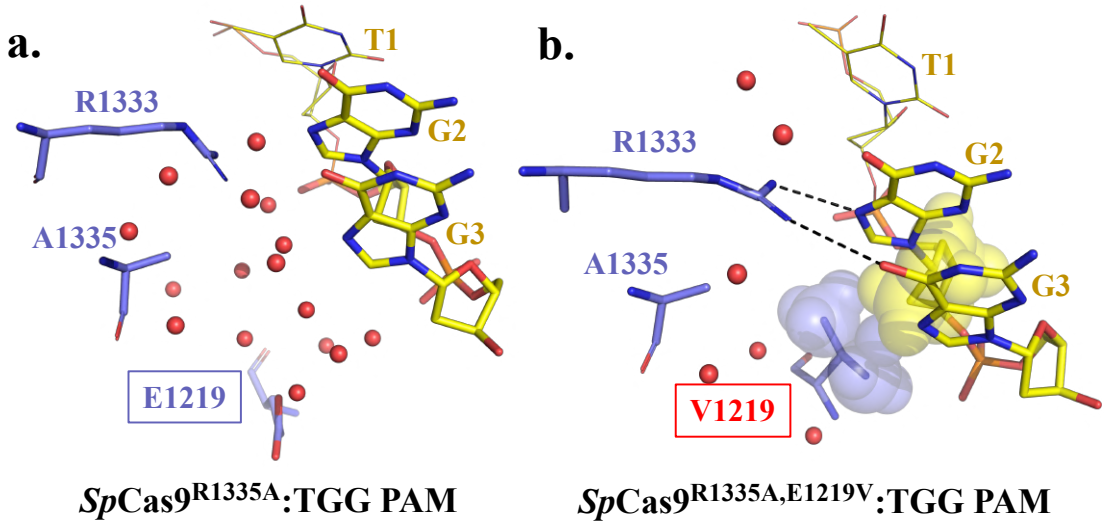


Figure 2.5. Structural comparison of 5'-TGG-3' PAM binding pocket in the pre-catalytic state of *Sp*Cas9^{R1335A} (a. left) and *Sp*Cas9^{R1335A,E1219V} (b. right). Key residues are depicted in sticks, the interaction network is represented by dotted lines. Hydrogen atoms are excluded for clarity. Water molecules around 4 Å of the PAM interacting region (N1 and O6 atoms of guanine, R1333 nitrogen atoms, and E1219/V1219) are shown as red spheres. The hydrophobic pocket created by V1219 and the deoxyribose sugar of 2nd guanine are highlighted with transparent sphere representation.

Recently, X-ray structure of the Cas9-NG variant of *Sp*Cas9 in complex with TGG-PAM has been reported (PDB 6AI6) (Nishimasu et al., 2018). Cas9-NG contains seven mutations (R1335V, E1219F, L1111R, D1135V, G1218R, A1322R, T1337R) Despite the significant difference between Cas9-NG (seven mutations, X-ray) and *Sp*Cas9^{R1335A,E1219V} (double mutation, MD), both the structures share common features (**Figure 2.6**), viz., the hydrogen bonding interaction between R1333 and G2, and the presence of hydrophobic pocket around the ribose sugar of G2 (Nishimasu et al., 2018). However, the intricate protein: DNA interactions in the X-ray and MD structures differ by (1) R1333 forms double hydrogen bonds with the G2 in Cas9-NG instead of H-bonds with both G2, G3 (in *Sp*Cas9^{R1335A,E1219V}), (2) T1337R mutation in the Cas9-NG, facilitated the hydrogen bonding with the G3 which is absent in the MD structure of the double mutant.

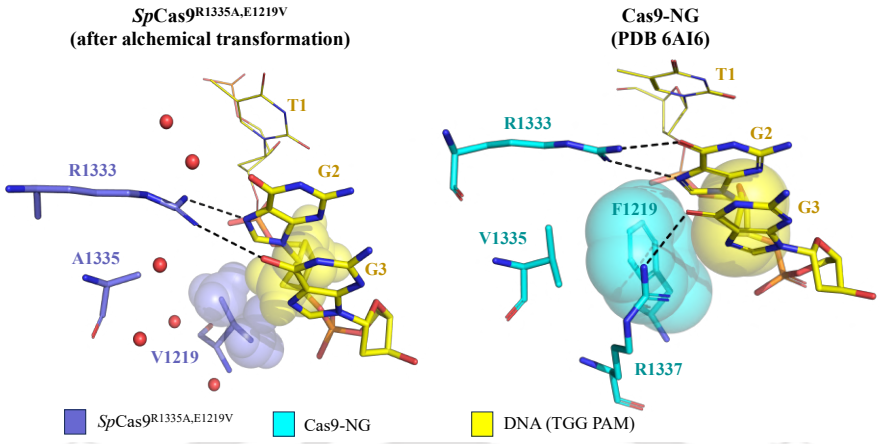


Figure 2.6. Comparison of MD structure of *Sp*Cas9^{R1335A,E1219V}:TGG (left) and X-ray structure of Cas9-NG:TGG (PDB 6AI6, right).

2.4. Discussion

Stringent PAM recognition by the CRISPR/Cas9 immune system allows distinction between foreign and self-DNA, but the same limits its use to genome editing applications by restricting the targetable sequences. Expansion of PAM readability can target wide DNA substrates and control metabolic rates and cell growth (Kim et al., 2020). The most-studied system for altered PAM readability is *Sp*Cas9 nuclease owing to its simple three-letter cognate PAM sequence (5'-NGG-3') and robust activity (Jiang & Doudna, 2017; Rath et al., 2015). Random mutagenesis of the PID of *Sp*Cas9 (E1219V, R1335A) resulted in variants with shifted PAM consensus (Guo et al., 2019; Nishimasu et al., 2018a). This work quantified the effects of E1219V mutation on the energetics of *Sp*Cas9 and *Sp*Cas9^{R1335A} binding to various DNA substrates differing in the PAM sequence (5'-TGG/TGT/GAT/TTG/TTT-3') using molecular dynamics free energy simulations as summarized in **Figure 2.2b**. X-ray structure of pre-catalytic WT *Sp*Cas9 in complex with its cognate PAM revealed specific hydrogen bonds between two arginine's of PAM interacting domain (PID, R1333, and E1219-salt-bridged R1335) and G2, G3 of the 5'-NGG-3' (**Figure 2.1**). Biochemical studies showed that the E1219V mutation in *Sp*Cas9 expands the PAM readability (Guo et al., 2019). It was believed that the E1219V mutation induced unrestricted rotamerization of R1335 and allowed *Sp*Cas9 to recognize multiple PAM sequences. The binding affinity of the

*Sp*Cas9 to canonical 5'-TGG-3' PAM is found to be independent of E1219→V1219 mutation (**Figure 2.2b**), supported by the conserved hydrogen bonding between PID and G2, G3 of 5'-TGG-3' (**Figure 2.3a,b**). Biochemical experiments showed that the wild-type and E1219V mutant of *Sp*Cas9 displayed similar cleavage activity for the substrate containing TGG PAM (Guo et al., 2019), corroborating the estimated energetics. Thus, R1335 side-chain rotamerization in response to E1219V mutation in *Sp*Cas9 is found to play no significant role in binding to dsDNA containing cognate PAM (5'-TGG-3') sequence.

Loss of hydrogen bonding between PID of wild-type *Sp*Cas9 and non-canonical 5'-TGT/GAT-3' (**Figure 2.3c,e**) accounts for the experimentally observed weaker recognition of non-canonical PAM relative to the canonical 5'-TGG-3' sequence (Guo et al., 2019). Interestingly, the E1219→V1219 mutation in *Sp*Cas9 improves the binding affinity by ~ 4 and ~2 kcal/mol for the substrate containing 5'-TGT-3' and 5'-GAT-3', respectively. The simulations further revealed that the E1219V mutation excludes water molecules, particularly from the pocket that accommodates the 3rd nucleotide base of the PAM (T3; **Figure 2.3d,f**). V1219 seems to provide a hydrophobic cushion for the 5-CH3 group of the thymine, resulting in the improvement of the binding affinity for TGT and GAT PAM sequences. The distance between the T3 methyl group and V1219 indeed decreased relative to the wild-type E1219 (**appendix Figure A2.8**), supporting the hydrophobic stabilization by V1219. Moreover, the desolvation also encourages electrostatic interaction between the PID of *Sp*Cas9 and the TGT/GAT sequence (**Figure 2.3d,f**). Hence, the E1219V mutation in *Sp*Cas9 shields the PAM sequence from water, thereby amplifying the binding affinity for 5'-TGT/GAT-3' sequences. This is also in line with biochemical experiments, which suggested that the E1219V mutation in the *Sp*Cas9 mutant improved the DNA cleavage activity for the substrates containing TGT and GAT PAM sequences, the former being noticeable (Guo et al., 2019). The calculated DDG for different DNA substrates can be compared with the relative cleavage activity from biochemical experiments (Guo et al., 2019). However, cleavage activity includes both the binding and the chemical step. Alchemical simulations confirmed that the magnitude of DDG's follows the order: $DDG^{TGG} < DDG^{GAT} < DDG^{TGT}$. The sign and magnitude of the calculated DDG's indicated excellent correlation with the relative cleavage activity. In the case of *Sp*Cas9 bound to canonical TGG PAM, no change in binding affinity ($\Delta\Delta G \sim 0$ Kcal/mol)

upon E1219V mutation might be the reason both WT *Sp*Cas9 and *Sp*Cas9^{E1219V} demonstrate high cleavage efficiencies. Although E1219V mutation in *Sp*Cas9 prefers both TGT and GAT PAM binding, the preference is higher for *Sp*Cas9^{E1219V} bound to TGT PAM, compared to GAT, which may explain the observed higher cleavage efficiency of *Sp*Cas9^{E1219V} in DNA containing TGT PAM compared to GAT PAM (Guo et al., 2019). The DNA cleavage activity is obviously controlled by kinetics, particularly the activation barrier for the chemical step. However, the differential binding affinity (PAM binding to wild-type-*Sp*Cas9 versus mutant-*Sp*Cas9^{E1219V}, $\Delta\Delta G$) can fine-tune the Boltzmann population of precatalytic *Sp*Cas9:sgRNA:dsDNA complex and control the cleavage activity. Thus, PAM selectivity can serve as the initial checkpoint for the Cas9 activity. The binding of DNA to Cas9:sgRNA complex (i.e., PAM binding followed by R-loop formation) is reported to be the rate-limiting step, whereas the subsequent DNA cleavage step is rapid (Raper et al., 2018). Therefore, tuning the Boltzmann population of the *Sp*Cas9 pre-catalytic state will surely alter the rate of cleavage activity. Michaelis-Menten constant K_M is inversely related to the substrate binding affinity. Thus, favourable *Sp*Cas9:DNA binding in response to protein mutation can improve the catalytic efficiency by lowering the K_M value.

This proposed mechanism appears implausible because, in general, solvent exclusion from the recognition site is known to amplify the stringency for cognate substrate selection by disfavouring non-cognate substrates. The penalty for incorrect substrate accommodation in the biomolecule can be significantly amplified by not allowing water molecules to compensate for the missing bonding requirements in the substrate binding pocket (e.g., mRNA decoding, protein: DNA recognition, protein: RNA interactions) (Kumar et al., 2017; Satpati et al., 2014; Satpati & Åqvist, 2014; Shukla et al., 2020). On the contrary, shielding the PAM binding pocket of *Sp*Cas9 from water enforces electrostatic contact between the protein and the T-rich non-cognate PAM, resulting in the expansion of PAM readability by *Sp*Cas9. Interestingly, unlike TGT, the TTG PAM was found to be non-selective between *Sp*Cas9 and *Sp*Cas9^{E1219V} (**Figure 2.2b**). Thus, the position of thymine in the PAM sequence was found to be crucial for ensuring selectivity.

We showed that R1335A substitution in the *Sp*Cas9 disrupts the hydrogen bonding between R1333 and 2nd guanine of the 5'-TGG/TGT-3' sequence and allows entry of water molecules in

the PAM binding pocket (**Figure 2.5a, appendix Figure A2.7**). Thus, the *Sp*Cas9^{R1335A} variant is expected to display very weak recognition for all the PAM sequences (5'-TGG or TGT or GAT-3'). Previous in vitro cleavage activities indeed confirmed that wild-type *Sp*Cas9 can efficiently cleave the plasmid containing TGG target, but R1335A displayed almost no activity (Nishimasu et al., 2018). E1219→V1219 mutation in the *Sp*Cas9^{R1335A} variant improved the binding affinity to all the PAM sequences studied in this work (TGG/TGT/GAT/TTG/TTT) due to the same reasons: (1) desolvating the PAM binding pocket, and (2) strengthening the PID-PAM hydrogen bonding. Noticeably, E1219→V1219 mutation in the *Sp*Cas9^{R1335A} significantly improves 5'-TGG-3' binding affinity ($\Delta\Delta G \sim 3.5$ kcal/mol, **Figure 2.2b**). However, the same mutation has no effect on the energetics of *Sp*Cas9 binding to 5'-TGG-3'. Thus, we hypothesized that the double mutant *Sp*Cas9^{R1335A,E1219V} is likely to display higher cleavage against 5'-TGG-3' PAM containing dsDNA relative to the single mutant *Sp*Cas9^{R1335A}.

2.5. Conclusion

To our knowledge, this is the first report that revealed the energetics of broadened PAM readability in response to E1219V mutation in the wild-type and R1335A-mutant *Sp*Cas9. The ability of *Sp*Cas9 to read non-canonical T-rich PAM sequences largely resides in the solvent exclusion effect of E1219V mutation. Solvent exclusion not only encourages thymine-containing PAM sequence (particularly in the third nucleotide position) by providing a hydrophobic cushion but also enforces new protein: nucleotide interactions (base-specific and or non-base-specific), thereby weakening the PAM stringency of *Sp*Cas9. R1335A mutation disrupts direct interactions between the *Sp*Cas9 and various PAM sequences (5'-TGG-3' or 5'-TGT-3' or 5'-GAT-3' or 5'-TTG-3' or 5'-TTT-3'), resulting in loss of activity. We hypothesized improved nuclease activity for substrate containing 5'-TGG-3' PAM sequence upon E1219V mutation in the inactive *Sp*Cas9^{R1335A} variant. The PAM selection process is certainly controlled by kinetics, but the underlying thermodynamics explains the fidelity of *Sp*Cas9 and, thus, the experimental findings.

Chapter 3

Effect of Multiple *Sp*Cas9 Mutations (Cas9-NG Mutations) on the Energetics of PAM Recognition

This chapter is published in J. Chem. Inf. Model, 2025, 65, 7, 3628–3639

The energetic basis for the enhanced PAM (protospacer adjacent motif) readability in engineered Cas9-NG (a variant of Cas9 from *Streptococcus pyogenes* (*Sp*Cas9) with seven mutations: R1335V, E1219F, D1135V, L1111R, T1337R, G1218R, and A1322R) remains a fundamental unsolved problem. While experimental mutagenesis highlights combinations that broaden PAM recognition, the energetic rationale for why certain mutations succeed or fail has been lacking. This chapter examines the changes in PAM (TGG, TGA, TGT, or TGC) binding affinity (ΔΔG) associated with each of the seven mutations in *Sp*Cas9 through rigorous alchemical simulations (sampling ~ 53 μs). The underlying thermodynamics (ΔΔG) account for the experimentally observed differences in DNA cleavage activity between *Sp*Cas9 and Cas9-NG across various DNA substrates. The interaction energies between *Sp*Cas9 and DNA are significantly influenced by the type and location of the amino acid mutations. Notably, the R1335V mutation disfavours DNA binding by disrupting critical interactions with the PAM. However, the destabilizing effect of R1335V mutation is mitigated by four advantageous mutations (E1219F, D1135V, L1111R, and T1337R), which primarily introduce non-base-specific interactions and enhance PAM readability. The hydrophobic substitutions (E1219F and D1135V) are particularly impactful, as they exclude solvent from the PAM binding pocket, strengthening electrostatic interactions in the low dielectric medium and increasing the stability of the non-cognate PAM complexes by ~ 2-5 kcal/mol. Additionally, L1111R and T1337R facilitate DNA binding by forming direct electrostatic contacts. In contrast, the charge mutations G1218R and A1322R do not effectively promote interactions with the negatively charged DNA, clearly demonstrating that the location of mutations is crucial

in shaping this interaction energetics. We demonstrated that stabilization of the Cas9-NG: non-cognate PAM complexes enable broader PAM recognition. This is primarily achieved through two mechanisms: (1) the establishment of new non-base-specific interactions between the protein and nucleotides and (2) the enhancement of electrostatic interactions within a relatively dry and hydrophobic pocket. The findings revealed that mutation-induced desolvation can improve the recognition of non-cognate PAMs, paving the way for the rational and innovative design of *Sp*Cas9 mutants.

This chapter examines how individual and combinatorial mutations within the PAM-interacting domain alter local solvation, residue flexibility, and base-specific contacts. We map the energetic landscape that enables broadened PAM profiles and clarify the distinct functional roles of key residues.

3.1. Background

CRISPR/Cas9 is a genome-editing tool discovered in prokaryotes as an immune system (Adli, 2018; Rath et al., 2015). Cas9, a multi-domain RNA-guided endonuclease, uses a single guide RNA (sgRNA) to bind to specific DNA sequences, introducing site-specific double-strand breaks for targeted editing (Jiang & Doudna, 2017; Jinek et al., 2012). Successful cleavage requires a specific Protospacer Adjacent Motif (PAM) and 20 nt base-complementarity between sgRNA and target DNA (Anders et al., 2014; Jinek et al., 2012). PAM recognition is crucial for Cas9 function, which triggers the local unwinding of the target DNA and enables base pairing, ensuring site-specific cleavage of viral DNA while preventing host-DNA cleavage (Anders et al., 2014; Nishimasu et al., 2014; Sternberg et al., 2014). The most widely used Cas9, derived from *Streptococcus pyogenes* (*Sp*Cas9), recognizes the 5'-NGG-3' PAM sequence (where N = any nucleotide) (Jiang & Doudna, 2017; Jinek et al., 2012). In *Sp*Cas9, base-specific hydrogen bonds between two arginines (R1333, R1335 of the PAM interacting (PI) domain of *Sp*Cas9) and two guanines of PAM (G2G3 of the DNA) ensure the PAM specificity (Anders et al., 2014; Nishimasu et al., 2014) (**Figure 3.1**).

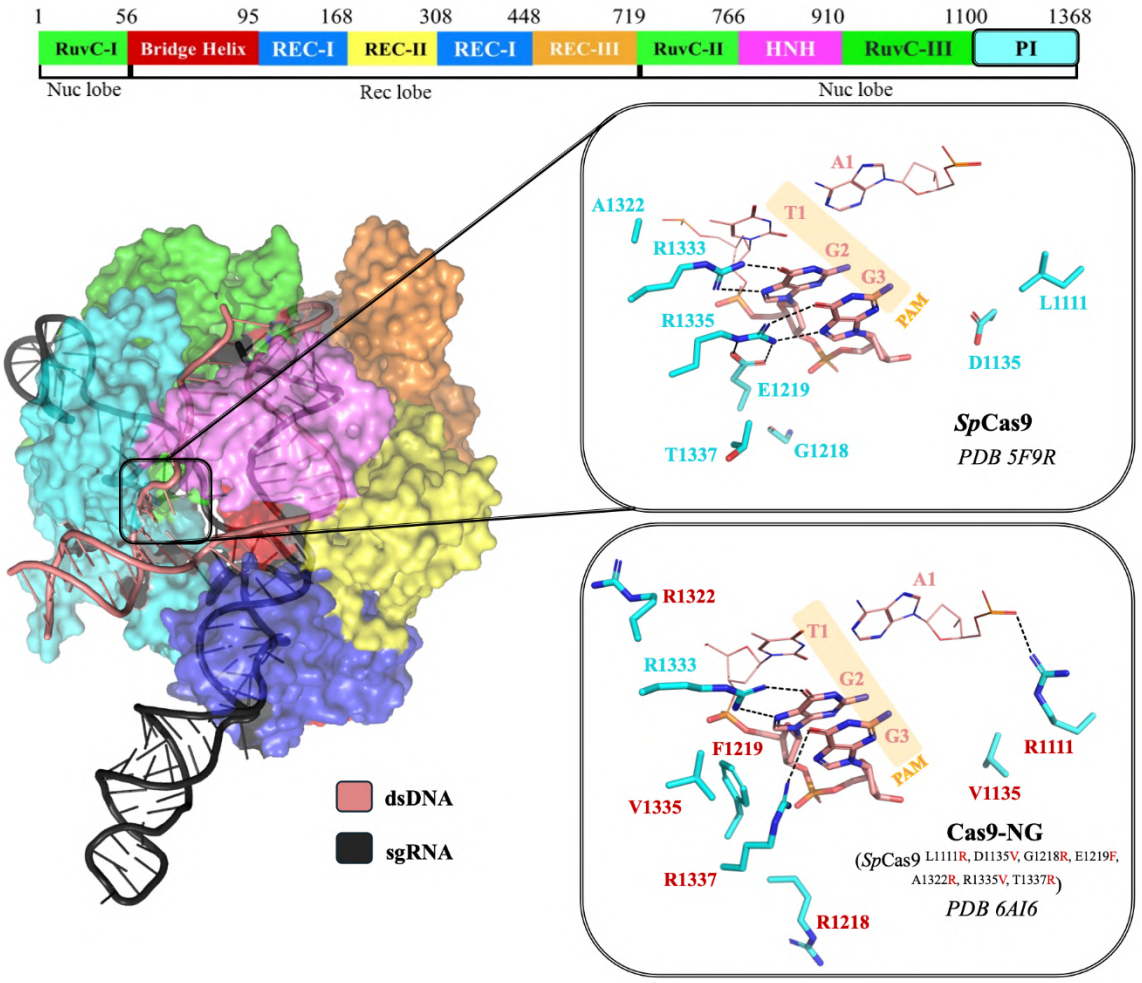


Figure 3.1. A colour-coded bar illustrating the domain organisation of *Sp*Cas9. The 3D structure of wild-type *Sp*Cas9:sgRNA: DNA (pre-catalytic complex), containing nucleic acids (ribbon) and domain organisation (coloured semi-transparent surface). A zoomed-in view (curved box) of the PAM binding pocket in *Sp*Cas9 and Cas9-NG (engineered Cas9 variant, lower right corner) is shown on the right. PAM (5'-TGG-3', yellow highlighted) and the residues of the "PI" domain (cyan) are numbered and shown in the sticks. Dotted lines depict interaction networks. Hydrogens are not shown for clarity.

The PAM stringency (viz., 5'-NGG-3' for *Sp*Cas9) is vital for achieving accurate genome editing; however, it also narrows the range of possible target sites for editing since the frequency of NGG in DNA of interest is limited (Hsu et al., 2014). Therefore, expanding the PAM readability by

Cas9 enzymes (Guo et al., 2019; Kleinstiver et al., 2015) is of great interest in advancing the technology. Engineered Cas9 variants (viz., VQR (Kleinstiver et al., 2015), VRER (Kleinstiver et al., 2015), SpG (Walton et al., 2020), SpRY (Walton et al., 2020), xCas9 (Hu et al., 2018), and Cas9-NG (Nishimasu et al., 2018; Ren et al., 2019) mutants of Cas9 protein) have been shown to have expanded PAM readability. Cas9-NG is a variant of the Cas9 protein that recognizes a 5'-NGN-3' PAM sequence (lifting the stringency of the third nucleotide) as opposed to 5'-NGG-3' (recognized by wild-type *Sp*Cas9) (Nishimasu et al., 2018; Ren et al., 2019). Expanded PAM recognition allows greater flexibility in designing guide RNAs, enabling researchers to target a wider array of genomic sites (Nishimasu et al., 2018; Ren et al., 2019). Cas9-NG contains seven mutations in the PAM interacting domain (L1111R, D1135V, A1322R, G1218R, E1219F, R1335V, T1337R) and was shown to be active against 5'-NGA or NGC or NGT-3' PAM targets (Nishimasu et al., 2018; Ren et al., 2019). However, the Cas9 mutants are produced in somewhat arbitrary ways. The rational design of new Cas9 mutants requires an understanding of the mechanisms behind PAM recognition, particularly in relation to their structures and the link to energetics, which represents a current knowledge gap.

This chapter examines the effects of seven Cas9-NG mutations (R1335V, E1219F, D1135V, L1111R, T1337R, G1218R, and A1322R) on the energetics of binding to double-stranded DNA (dsDNA) with different PAM sequences (5'-TGG or TGA or TGT or TGC-3'). Using molecular dynamics free energy calculations, we show that the difference in DNA binding affinity ($\Delta\Delta G$) varies significantly based on the type and location of the mutation. Our results clarify the differences in cleavage activity (Nishimasu et al., 2018) between Cas9-NG and wild-type *Sp*Cas9. The mutation in R1335V disrupts *Sp*Cas9:PAM interactions, while four mutations (E1219F, D1135V, L1111R, and T1337R) enhance these interactions by forming non-base-specific contacts and widening PAM recognition. In contrast, G1218R and A1322R do not affect DNA binding. The findings establish a connection between the atomic structure, energetics, and cleavage activity related to PAM selectivity by *Sp*Cas9.

3.2. Methodology

3.2.1. Molecular Dynamics Setup

The X-ray structure (PDB 5F9R (Jiang et al., 2016)) of the precatalytic *Sp*Cas9 in complex with sgRNA and dsDNA (containing canonical 5'-TGG-3' PAM sequence) was retrieved from Protein Data Bank (Berman et al., 2000). A spherically truncated system of radius 25 Å, centered at the mutating residue (L1111R or D1135V or A1322R or G1218R or E1219F or R1335V or T1337R), was extracted from the template X-ray structure (**appendix Table A3.1**). The robustness of the structure, as well as the energetics obtained from the truncated system, was further examined by repeating the calculations for the full system (*Sp*Cas9:TGG and *Sp*Cas9^{R1335V}: TGG), ensuring convergence (**appendix Table A3.2**). A larger truncated spherical system (~35 Å radius) centered at the G2 position of the PAM sequences (TGG, TGA, TGT, TGC) was employed for models used in calculating the simultaneous effect of all seven mutations (L1111R, D1135V, A1322R, G1218R, E1219F, R1335V, T1337R). Models of *Sp*Cas9 bound to non-canonical PAM sequences (5'-TGA-3'or 5'-TGT-3' or 5'-TGC-3' PAM) were generated by appropriate substitutions in the dsDNA using PyMol. The list of all the structural models used from MD simulations is provided in **appendix Table A3.1**. Protein structure files (PSF) for individual models were generated with the help of the CHARMM-GUI server (Jo et al., 2008; Park et al., 2023). The standard CHARMM36 force field was used to describe the biomolecules (Huang & Mackerell, 2013; MacKerell et al., 1998), and the water molecules were described by TIP3P model (Jorgensen et al., 1998). Keeping the biomolecule at the centre, an explicit equilibrated water box was overlaid. Monovalent counterions have been added to the solvated box to ensure charge neutrality. The resulting system was subjected to energy minimization, equilibration, and production dynamics (the detailed protocol explained in chapter 1, (**section 1.8.2**). Simulations were conducted at a temperature of 310 K and a constant pressure of 1 bar. Langevin Dynamics (Phillips et al., 2005) and a Langevin Piston (Feller et al., 1995; Martyna et al., 1994) were used to control temperature and pressure, respectively. The periodic boundary conditions were employed throughout the simulations, with a time step of 2 fs. A cut-off distance of 12 Å was used to truncate Van der Waals interactions. The long-range electrostatic interactions were calculated utilizing the particle mesh Ewald method (Darden et al., 1993) with tinfoil boundary conditions (Bogusz et al., 1998;

Hummer et al., 1996). The use of tinfoil boundary conditions is particularly important for maintaining charge neutrality during alchemical transformations (see next section). The tinfoil boundary conditions ensure that the gain/loss of a positive/negative charge along the alchemical pathway is implicitly neutralized by a compensating charge density that automatically spreads uniformly throughout the simulation box, maintaining a net zero electrostatic potential (Levrel & Maggs, 2008; Lin et al., 2014; Roberts & Schnitker, 1995; Rocklin et al., 2013; Simonson & Satpati, 2013). Thus, the overall system remains charge-neutral without requiring manual adjustment of counterions.

The simulations were run in NAMD 3.0 (H. Chen et al., 2020; Phillips et al., 2005) software. The trajectories (snapshots separated by ten ps intervals) obtained from production dynamics were used for data collection for structural analysis. To ensure adequate sampling, several independent replicas (3 for each model) of simulations were carried out by considering different initial velocity distributions, resulting in a total of 27.4 μs of trajectory considered for analysis. The structures obtained after production dynamics from various independent trajectories were subjected to alchemical transformation to estimate the relative binding affinity.

3.2.2. Relative Binding Free Energy Estimations Using Alchemical Simulations

Single amino-acid alchemical transformation in *Sp*Cas9 (R1335 → V1335, E1219→F1219, D1135→V1135, L1111→R1111, T1337→R1337, G1218→R1218, and A1322→R1322) was carried out on the final structures obtained from the conventional molecular dynamics simulations. The change in DNA binding free energy (ΔΔG) of *Sp*Cas9 in response to alchemical transformation was estimated by employing an appropriate thermodynamic cycle (**Figure 3.2a**). Vertical aems signify DNA binding, whereas horizontal legs correspond to the alchemical transformation of *Sp*Cas9 to *Sp*Cas9^{Mutant} in dsDNA's presence (upper arm) and absence (lower arm). Free Energy Perturbation (FEP) method (Zwanzig, 2004) was employed to compute the alchemical free energies (ΔG_{comp} and ΔG_{free}) associated with the horizontal arms (**Figure 3.2a**), as elaborated in chapter 1 (**section 1.8.3**).

Alchemical transformations for single amino acid mutations were performed using 51 λ windows spaced from 0 to 1. Each window was simulated for 3–5 ns depending on system size, with the first 1 ns discarded as equilibration. For simultaneous (all seven) alchemical transformations a greater number of λ windows (201 equally spaced λ points between λ=0 to λ=1) and a larger truncated system (~ 35 Å radius centred at the N3 atom of G2 base). Needless to say, the alchemical transformation of multiple residues demands increased stratifications and sampling to ensure accurate free energy estimation (Pohorille et al., 2010). Thus, we increased the stratification to 201 windows for multiple-alchemical transformations. The free energy differences and the associated statistical errors were calculated using the ParseFEP plugin of VMD (Humphrey et al., 1996; Liu et al., 2012). Good convergence of the calculated free energies was evident from three independent replicas (ΔG differs by < 1 kcal/mol, **appendix Table A3.2**) and a reasonable statistical error (< 0.5 kcal/mol, **appendix Table A3.2**). Convergence was also monitored by comparing ΔG (two neighbouring windows) as a function of time, and the free-energy profiles plateaued consistently within the final third of the trajectories (**appendix Figure A3.1**). The averaged ΔG values (from all three trials) were reported as ΔG_{comp} or $\Delta G_{free,}$ and the associated errors are standard errors of the mean (s.e.m). The relative binding free energy ΔΔG was calculated as $\Delta\Delta G = \Delta G_{comp} - \Delta G_{free,}$ and the errors for ΔΔG were obtained by propagating the errors associated with ΔG_{comp} and ΔG_{free}. The sign of ΔΔG (negative/positive) suggests that the DNA binding is favoured/disfavoured in response to *Sp*Cas9 mutation. The magnitude of the calculated ΔΔG is attributed to the strength of the preference. These calculations were repeated for various DNA differing in their PAM sequence (cognate: TGG, non-cognate: TGA or TGT or TGC), thus estimated the effect of *Sp*Cas9 mutation on the binding to different PAM-containing DNAs. To ensure adequate sampling and good convergence, a total of 25.3 μs of simulation was performed for alchemical free energy calculations.

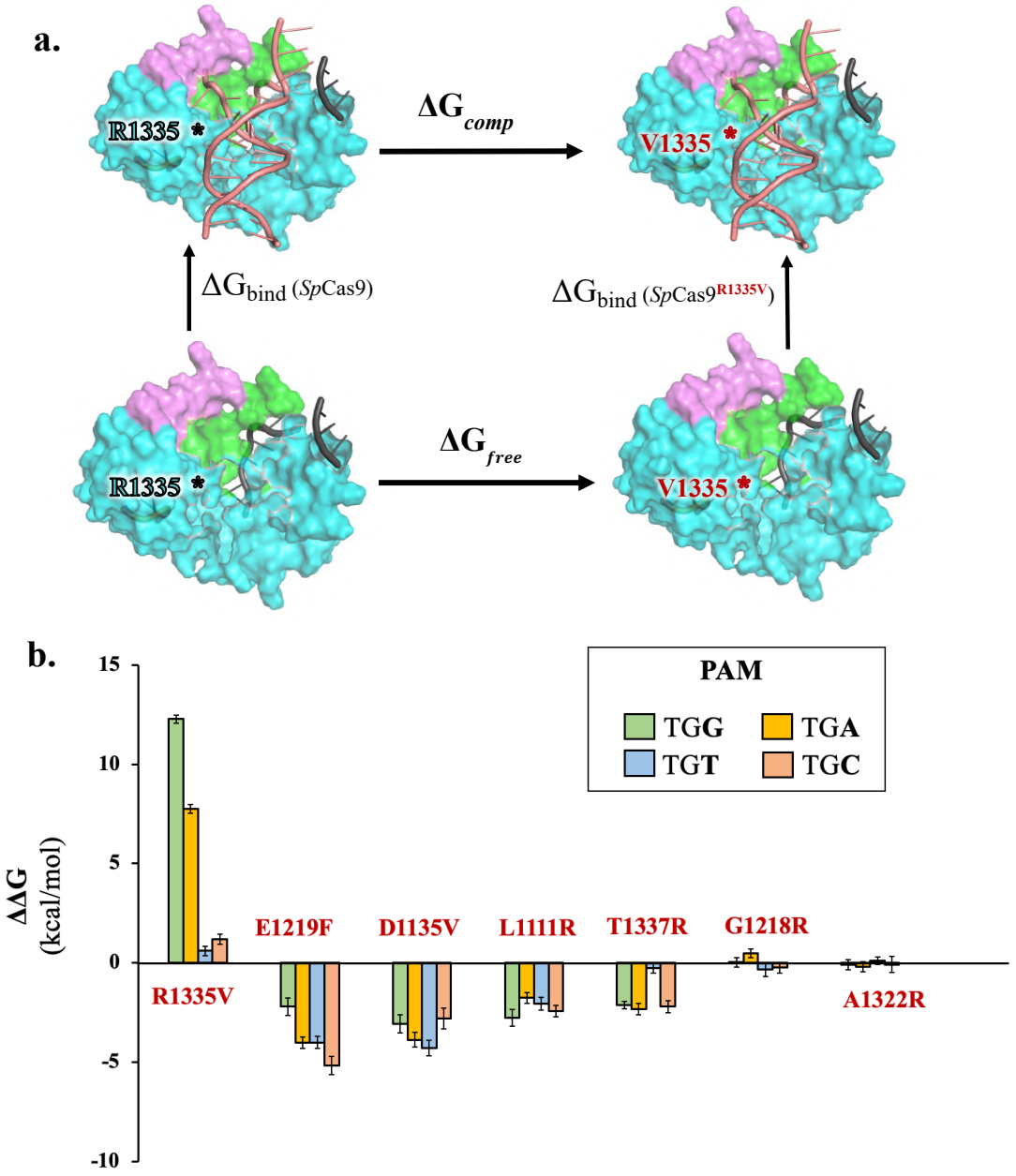


Figure 3.2. (a) Thermodynamic cycle and the effect of *Sp*Cas9 mutation on binding affinity to dsDNA. Vertical legs represent binding, while horizontal legs depict the alchemical transformation of the side chain from R1335 (wild-type) to V1335 in the *Sp*Cas9, either in complex with dsDNA (above) or free in water (below). The free energy difference between wild-type *Sp*Cas9 and mutant *Sp*Cas9^{R1335V} binding to dsDNA is $\Delta\Delta G = \Delta G_{comp} - \Delta G_{free} = \Delta G_{bind}$ (*Sp*Cas9$^{R1335\rightarrow V1335}$) - ΔG_{bind}(*Sp*Cas9). (b) Calculated binding free energy differences ($\Delta\Delta G$) for wild-type versus mutant *Sp*Cas9 binding to various dsDNA, each with a different PAM sequence (5'-TGG, TGA, TGT, or TGC; color-coded bars). Error bars, 1 s.e.m.

3.3. Results

3.3.1. Structural and Mechanistic Views of PAM Recognition by *Sp*Cas9

X-ray structures (PDB 5F9R and 4UN3) (Anders et al., 2014; Jiang et al., 2016) indicate that the cognate protospacer adjacent motif (PAM), specifically 5'-NGG-3', is recognized by two arginine residues (R1333 and R1335) located within the PI domain of wild-type *Sp*Cas9 (**Figure 3.1**). The presence of four base-specific hydrogen bonds between these arginine residues and the Hoogsteen edge of the G2G3 region of the cognate PAM, as illustrated in **Figure 3.1**. A strong intrinsic preference for arginine for guanine bases has been well documented (Helene, 1977; Hossain et al., 2023, 2025). Arginine's guanidinium group is known to form optimal hydrogen bonds with the O6 and N7 atoms of guanine. Therefore, it was hypothesized that wild-type *Sp*Cas9 disfavours noncognate PAM targets by disrupting interactions between arginine and the PAM when guanine is substituted with other bases like adenine/thymine/cytosine, thereby ensuring selectivity. Indeed, simulations of non-cognate complexes showed the loss of base-specific interactions (**appendix Figure A3.2**) between the non-cognate PAM (TGA or TGT or TGC) and wild-type *Sp*Cas9. The substitution of non-cognate DNA (5'-TGA, TGT, or TGC-3') in the pre-catalytic complex weakens the interaction between R1335 and the third base (A3, T3, or C3) due to obvious electrostatic repulsion (see **appendix Figure A3.2** and **appendix Tables A3.3** and **A3.4**). This is further supported by the increased flexibility of R1335 (**appendix Figure A3.3b**). A single water-mediated interaction was evident between T3(O4)/A3(N7) and R1335 (**appendix Figure A3.2b, c**), while there were no interactions between C3 and R1335 (**appendix Figure A3.2d**). Notably, the MD structure of *Sp*Cas9:TGC demonstrates that the presence of C3 can also disrupt the interaction in the 2nd base (R1333:G2 is lost), supported by the increased flexibility of R1333 (**appendix Figure A3.3a**) and water exposure of the PAM binding pocket (**appendix Figure A3.2d**). It can be argued that the strong electrostatic repulsion between the amino group of C3 and the positively charged guanidium tip of arginine of the protein disrupted the interactions between arginine and DNA. A previous study (Hossain et al., 2023) found that positively charged lysine or arginine in the major groove of cytosine is highly unfavourable, leading to cytosine recognition primarily through direct interactions with negatively charged aspartate or glutamate

residues in the protein. We believe that the electrostatic repulsion between R1335 and C3 facilitates solvent entry, which is crucial for disrupting the interactions between R1333 and DNA (**appendix Figure A3.2d**).

The PAM-interacting domain (PI) of wild-type *Sp*Cas9 recognizes the cognate 5'-TGG-3' PAM sequence in the non-target strand, triggering DNA unwinding and forming the precatalytic complex through the base pairing of sgRNA with target DNA (**Figure 3.1**). Next, *Sp*Cas9 cleaves double-stranded DNA (dsDNA), specifically three nucleotides upstream of the PAM sequence, by triggering a conformational change in the HNH and RuvC nuclease domains (Jinek et al., 2012; Palermo, Miao, et al., 2017; Wang et al., 2023). HNH and RuvC domain involves Mg^{2+} ions to cleave the target and non-target strand of DNA (Das et al., 2023; Jinek et al., 2012; Nierzwicki et al., 2022). Indeed, previous studies have demonstrated that PAM binding allosterically triggers long-range conformational changes and cross-talk between nuclease domains (RuvC and HNH) (Palermo et al., 2017). Thus, both the stability of the pre-catalytic complex and the conformational change (nuclease domains) are likely to contribute to the discriminatory power of the sgRNA-programmed *Sp*Cas9, although the effect has never been quantified.

However, the binding of dsDNA to the Cas9:sgRNA complex is reported to be slow (rate-limiting), while the subsequent step of DNA cleavage occurs rapidly (Raper et al., 2018). Thus, the stability of the pre-catalytic complex seems to be the key to ensuring the accuracy of *Sp*Cas9 in genome editing. Experiments showed that several mutations (viz., E1219V, R1335A, D1135E, D1135V, T1337R) in the *Sp*Ca9 can significantly broaden the PAM readability and affect the DNA cleavage activity (Guo et al., 2019; Kleinstiver et al., 2015; Nishimasu et al., 2018). In Chapter 2, we showed that the E1219V mutation in *Sp*Cas9 alters the stability of the pre-catalytic complex (containing cognate or non-cognate PAM sequence) and the calculated energetics corroborating the experimental cleavage activity. An excellent correlation between binding affinity and cleavage activity suggested that the cleavage step is controlled by the PAM recognition, and the underlying thermodynamics seem sufficient to explain the experimental cleavage activity of *Sp*Cas9 (for DNA containing TGG or GAT or TGT PAM).

An engineered variant of *Sp*Cas9 (Cas9-NG, containing seven mutations: L1111R, D1135V, A1322R, G1218R, E1219F, R1335V, and T1337R) displays expanded PAM readability

(recognize NG PAM) (Nishimasu et al., 2018; Ren et al., 2019). Moreover, the X-ray structure of the Cas9-NG: TGG complex (PDB 6AI6 (Nishimasu et al., 2018)) reveals that new non-base specific interactions stabilize the complex (Nishimasu et al., 2018). Note the mutations have been introduced in a more or less random fashion to generate Cas9-NG. The effect of each mutation on the stability of the complex has never been quantified. The underlying energetics for expansion or stringency of PAM recognition by Cas9 is unknown.

3.3.2. Effect of *Sp*Cas9 mutation on DNA binding affinity

To quantify the energetics and elucidate the contributions of specific amino acid mutations (R1335V, E1219F, D1135V, L1111R, T1337R, G1218R, and A1322R) in PAM recognition, encompassing both cognate and non-cognate targets, we performed extensive molecular dynamics simulations (sampling ~ 52.7 μs). These simulations involved both wild-type and mutant *Sp*Cas9 variants in pre-catalytic complexes. The calculations involve computing the change in binding affinity upon *Sp*Cas9 mutation (viz., R1335→V1335, Figure 2) for a specific PAM sequence (viz., TGG) in the pre-catalytic complex. An appropriate thermodynamic cycle was employed (**Figure 3.2a**), and the relative binding free energy difference (ΔΔG) was estimated using alchemical free energy calculations (**Figure 3.2b**). We repeated the simulations for seven individual mutations (R1335V, E1219F, D1135V, L1111R, T1337R, G1218R, and A1322R) across different pre-catalytic complexes that had different PAM sequences: TGG, TGA, TGT, or TGC. Additionally, the simultaneous impact of all seven simultaneous mutations was quantified (**appendix Table A3.2**).

The difference in binding free energy between *Sp*Cas9 and the *Sp*Cas9$^{\text{single-mutant}}$ for various PAM sequences (5'-TGG, TGA, TGT, or TGC-3') is illustrated in **Figure 3.2b**. The calculated energetics (ΔΔG, shown in **Figure 3.2b**) highlight several notable features. First, the R1335V mutation is strongly disfavoured in the complex when the DNA contains the 5'-TGG or 5'-TGA-3' PAM sequences. In contrast, there is no significant energetic penalty associated with the 5-TGT or 5'-TGC-3' PAM sequences. Notably, when a purine base is present at the 3rd position of the PAM (G3 or A3), it enhances selectivity by significantly disfavouring the *Sp*Cas9$^{\text{R1335V}}$ compared to wild-type *Sp*Cas9 (by more than 12 or 7 kcal/mol). Conversely, the presence of a pyrimidine

base in the 3rd position of the PAM (either T3 or C3) shows weak discrimination, disfavouring the *Sp*Cas9^{R1335V} by only ~ 1 kcal/mol relative to wild-type *Sp*Cas9. Second, regardless of whether the PAM sequence is cognate or non-cognate, the *Sp*Cas9 mutations (E1219F, D1335V, T1337R, and L1111R) are generally favoured within the complex (by ~ 2-5 kcal/mol). The exception is the T1337R mutation, which does not exhibit selectivity in the 5'-TGT-3' complex. Third, the stability of the complex is more or less unaffected (less than 0.5 kcal/mol) by the substitutions G1218R and A1322R in *Sp*Cas9, and the stability remains independent of the nature of the PAM sequence.

3.3.3. The link between Calculated Energetics and Structures

Molecular dynamics simulations show that the variation in the number of hydrogen bonds between wild-type and mutant complexes directly correlates with their relative stability, which accounts for their differences in PAM discrimination. The strongest discrimination against the R1335→V1335 mutation in the *Sp*Cas9:TGG complex (ΔΔG ~ +12 kcal/mol, **Figure 3.2b**) arises from the loss of four hydrogen bond interactions between the G2G3 and arginine residues (R1333, R1335, **Figure 3.3b**). This loss leads to increased solvent exposure (**appendix Figure A3.4a**), increased flexibility of R1333 (**appendix Figure A3.3a**), and loss of residue correlation in the PAM binding pocket (**appendix Figure A3.5**) as a compensatory mechanism for the missing hydrogen bond interactions (**Figure 3.3a,b**; **appendix Table A3.3**). The R1335→V1335 mutation in the *Sp*Cas9:TGA complex incurs a penalty of over +7 kcal/mol due to the loss of two crucial hydrogen bonds (as shown in R1333…G2, **Figure 3.3c,d**, **appendix Table A3.3**). No hydrogen bonding interactions between the TGC PAM and arginine residues (R1333, R1335) are observed in the case of *Sp*Cas9:TGC and *Sp*Cas9^{R1335V}:TGC complexes (**Figure 3.3g,h**). Thus, no change in protein: nucleotide interaction network in response to R1335V mutation could explain the weak discrimination for the TGC target (ΔΔG ~1 kcal/mol, **Figure 3.2b**). The *Sp*Cas9:TGT complex is stabilized by two key factors: (1) two hydrogen bonds between R1335 and G2, and (2) a water-mediated interaction between R1335 and T3 (**Figure 3.3e**). It is noteworthy that the water-mediated interaction is disrupted when R1335 is mutated to V1335. Nevertheless, this mutation facilitates the development of a favorable hydrophobic interaction between the hydrophobic valine residue and the methyl group of T3 (**Figure 3.3f**). Thus, the net effect of R1335V mutation

(*Sp*Cas9) on TGT binding is negligible ΔΔG ~ 0. Interestingly, the hydrophobic contact between V1335 and T3 helps shield the R1333:G2 pair from the solvent (**Figure 3.3f; appendix Figure A3.4a**). The shielding effect stabilizes the R1333:G2 interaction (**Figure 3.3f**; **appendix Figure A3.4a**, **appendix Table A3.3**).

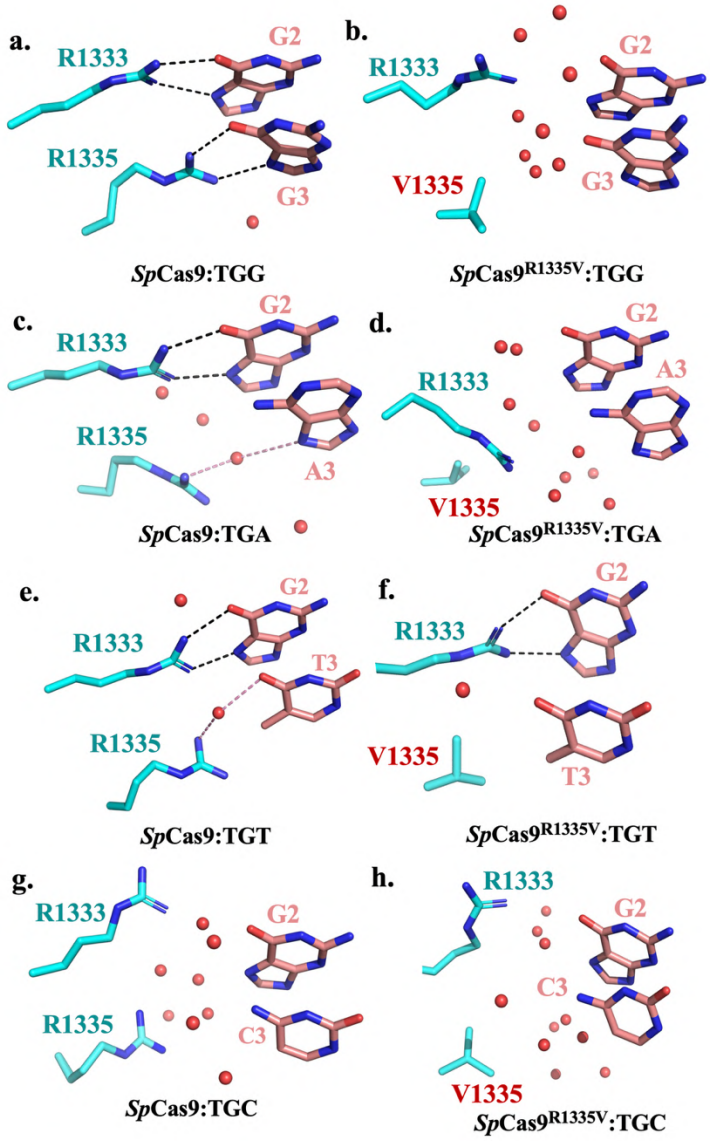


Figure 3.3. Zoomed in view of precatalytic *Sp*Cas9 and *Sp*Cas9^{R1335V} structures in complex with (a, b) TGG PAM, (c, d) TGA PAM, (e, f) TGT PAM, (g, h) TGC PAM. R1335→V1335 mutation resulted in loss of four and two hydrogen bonds for *Sp*Cas9:TGG and SpCas9:TGA respectively. The PAM and amino-acid residues are shown with pink and cyan colours, respectively, and water molecules within 4 Å of

Hoogsteen edges are depicted as red spheres. Amino acid main chains and hydrogen atoms are not shown due to clarity. Black dotted lines represent direct *Sp*Cas9:PAM interactions, while pink dotted lines represent water-mediated interactions. Water molecules are represented as red spheres. The same colour scheme is followed throughout the chapter.

Recent studies (Guo et al., 2019; Hossain et al., 2025) highlighted rigid conformation of R1335 sandwiched between G3 and E1219 residue in the cognate *Sp*Cas9:TGG complex, which is in line with our observations (**appendix Figure A3.3b**). This low entropy rigid conformation of R1335 ensures the high specificity of *Sp*Cas9 for NGG PAMs. It was claimed that mutating E1219 to neutral residues enhances the flexibility of R1335, enabling it to adopt alternative conformations for non-canonical PAM sequences (Guo et al., 2019; Hossain et al., 2025). We demonstrated that the effect of the E1219F mutation is twofold. First, it increases the flexibility of R1335 (**appendix Figure A3.3b**), as highlighted in previous studies (Guo et al., 2019; Hossain et al., 2025). Second, it excludes water molecules from the PAM binding pocket (**Figure 3.3; appendix Figure A3.4b**). This increases local hydrophobicity, significantly promoting new interactions between the protein and non-cognate DNA while also strengthening the existing local electrostatic interactions.

In the *Sp*Cas9:TGG complex, four hydrogen bonds between the protein and G2G3 are strengthened by the solvent-excluded low dielectric environment created by the E1219 → F1219 substitution (**Figures 3.4a,b appendix Figure A3.4b**). This favours the *Sp*Cas9^{E1219F}: TGG complex over the *Sp*Cas9:TGG complex, with a ΔΔG of approximately -2 kcal/mol (**Figure 3.2b**). In case of Non-canonical PAMs (TGA, TGT), the solvent exclusion of the PAM binding pocket by E1219F mutation promoted new salt-bridge interactions with DNA backbone (**Figure 3.4 d, f**), thus favouring TGA/TGT binding by ~4 kcal/mol (**Figure 3.2b**). Note that no direct protein-PAM interactions are observed in the *Sp*Cas9:TGC complex (**Figure 3.4 g**). Interestingly, the F1219 mutation in *Sp*Cas9:TGC introduced new protein-PAM interactions (base-specific R1333..G2 and salt-bridge R1335..G3, **Figure 3.4 g, appendix Table A3.4**). Thus, The E1219F mutation in the TGC target showed strongest preference with a ΔΔG of approximately -5 kcal/mol (**Figure 3.2b**). The desolvation of the PAM binding pocket due to the E1219F mutation (**appendix Figure A3.4b**) promoted the formation of a non-specific salt-bridge R1335...phosphate

interaction. This non-specific interaction favoured the binding of *Sp*Cas9^{E1219F} to the 5'-TGA or TGT-3' PAM compared to the wild-type *Sp*Cas9, resulting in a ΔΔG ~ - 4 kcal/mol (**Figure 3.2b**).

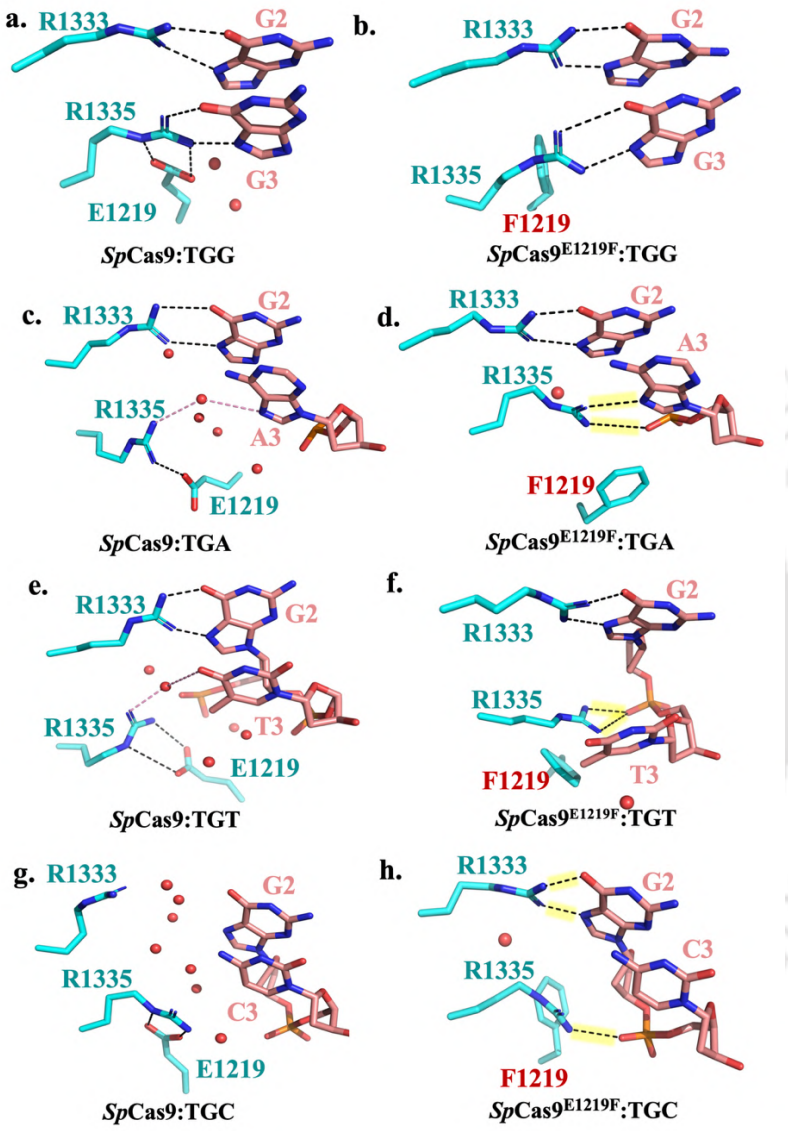


Figure 3.4. Zoomed in view of precatalytic *Sp*Cas9 and SpCas9^{E1219F} structures in complex with (a, b) TGG PAM, (c, d) TGA PAM, (e, f) TGT PAM, (g, h) TGC PAM. R1335→V1335 mutation resulted in loss of four and two hydrogen bonds for *Sp*Cas9:TGG and SpCas9:TGA respectively. Aromatic amino acid substitution, i.e., E1219 → F1219 (right-side) in the *Sp*Cas9 efficiently excludes solvent molecules (red sphere) from the PAM binding pocket. Thus, solvent excluded low dielectric medium boosts the existing electrostatic interaction and stabilizes the (a, b) *Sp*Cas9^{E1219F}:TGG complex. The low dielectric drastically

stabilizes the (d, f, h) *Sp*Cas9^{E1219F}:TGA/TGT/TGC complex by forming protein: DNA electrostatic contact (yellow highlighted).

No direct interaction is observed between D1135 of wild-type *Sp*Cas9, located in the minor groove of the PAM residues, and the DNA in molecular dynamics (MD) simulations (**Figures 3.5a** and **appendix Figure A3.6**) or the X-ray structure (**Figure 3.1**). The mutation of D1135 to V1335 in *Sp*Cas9 results in the solvent exclusion of the region (**appendix Figure A3.4c**), which facilitates the formation of a new polar interaction between the protein and the PAM, specifically involving S1136 and the ribose of G3 (**Figures 3.5b, appendix Figure A3.6** and **appendix Table A3.5**). Thus, it increases the preference of *Sp*Cas9^{D1135V} to bind to all four PAM sequences (by ΔΔG ~ -3 kcal/mol, **Figure 3.2b**). The stable polar interaction in the *Sp*Cas9^{D1135V} mutant reduced the flexibility of the S1136 residue (**appendix Figure A3.3c**). L1111→R1111 mutations in *Sp*Cas9 boost non-specific salt bridge interaction with the complementary strand of the PAM sequence by introducing a positive charge (**Figures 3.5 c, d**, and **appendix Figure A3.7**) and stabilize the complex by ~2.5 kcal/mol, irrespective of the PAM sequences.

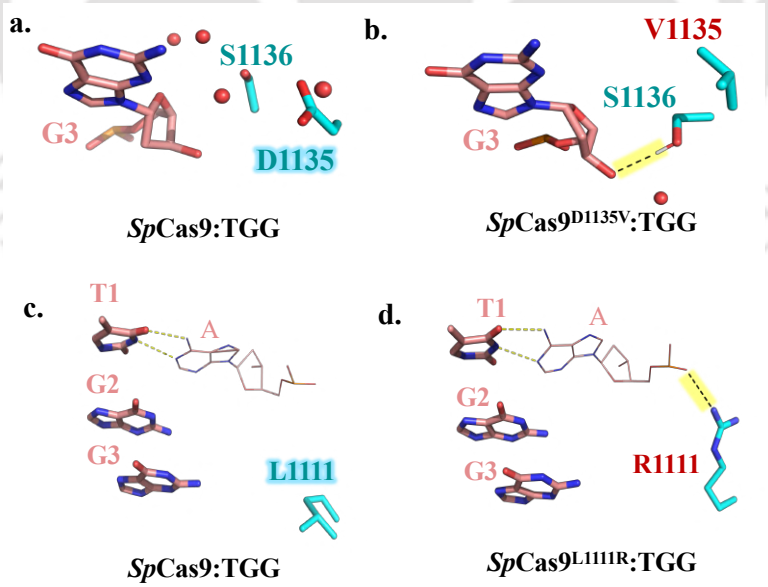


Figure 3.5. (a, b) D1135 (left) →V1135 (right) mutation in *Sp*Cas9:TGG complex desolvate the dsDNA (PAM) minor groove and drive a new non-specific polar interaction (sugar of DNA and serine side-chain, yellow highlighted) and stabilize the complex. (c, d) L1111→ R1111 mutation in *Sp*Cas9 promotes non-specific salt-

bridge interaction (yellow highlighted, R1111 side-chain and phosphate of the PAM complementary DNA strand, shown in lines). Water molecules are represented as red spheres. The DNA base pairing interaction network is shown on yellow dotted lines.

Similarly, T1337 → R1337 substitution in *Sp*Cas9 also induces a non-specific salt-bridge interaction with the DNA (**Figure 3.6 b, appendix Figure A3.8**; **appendix Table A3.6**) stabilizing the complex by ~ 2 kcal/mol, with the only exception in *Sp*Cas9^{T1337R}: TGT, where no new interactions are observed (**Figure 3.6 c**), resulting in very weak discrimination of less than ~ 0.25 kcal/mol (**Figure 3.2b**). The hydrophobic methyl group of T3 of TGT PAM repels the positively charged R1337 in the *Sp*Cas9^{T1337R}: TGT complex, preventing the formation of non-specific salt-bridge interactions (**Figure 3.6 c**). The MD structure of *Sp*Cas9^{T1337R}: TGG complex (**Figure 3.6b**) differs from the X-ray structure (Nishimasu et al., 2018) of Cas9-NG: TGG (**Figure 3.1**), where R1337 forms a salt-bridge with the DNA phosphate instead of base-specific interaction (R1337..G3, **Figure 3.1**) observed on the experimentally observed structure. X-ray structure of Cas9-NG: TGG complex (**Figure 3.1**) reveals that R1337 forms an H-bond with G3, compensating for the loss of interactions due to the presence of adjacent mutation R1335V. Moreover, in vitro DNA cleavage activity assay(Nishimasu et al., 2018) of wild-type *Sp*Cas9 and Cas9-NG confirmed significantly enhanced activity for the latter for non-cognate PAM sequences, including TGT, thus do not comply with the surprisingly low selectivity (ΔΔG ~ -0.25 kcal/mol, **Figure 3.2 b**). The observed deviation may be due to the very different nature of *Sp*Cas9^{T1337R} and *Sp*Cas9-NG, which differ by seven mutations, including T1337R. Analysis of the X-ray structure of Cas9-NG: TGG (**Figure 3.1**) indicated that two neighbouring hydrophobic residues of T1337 (i.e., V1335 and F1219) might influence the local environment, thus structure and energetics. Thus, the alchemical calculations (T1337→R1337) were repeated for the double mutant of *Sp*Cas9^{R1335V,E1219F} and the ΔΔG is reported in **Figure 3.6 a**. The effect of T1337→R1337 mutation in the double mutant *Sp*Cas9^{R1335V,E1219F} drastically boosts the PAM discrimination and favours DNA binding, including TGT PAM. Moreover, the MD structures *Sp*Cas9^{R1335V,E1219F,T1337R} could reproduce base-specific R1337..G3 interactions (**Figure 3.6 b**; **appendix Figure A3.9; appendix Table A3.7**) observed in the X-ray structure (**Figure 3.1**),

compensating the loss of interaction by R1335V mutation. Hence, it can be argued that a combination of three mutations (R1335V, E1219F, T1337R) in *Sp*Cas9 can replicate Cas9:PAM interactions in the Cas9-NG variant and corroborate the enhanced cleavage activity for TGT PAM. This highlights the cooperative effect among these three residues (V1335, F1219, and R1337) in improving PAM binding affinity.

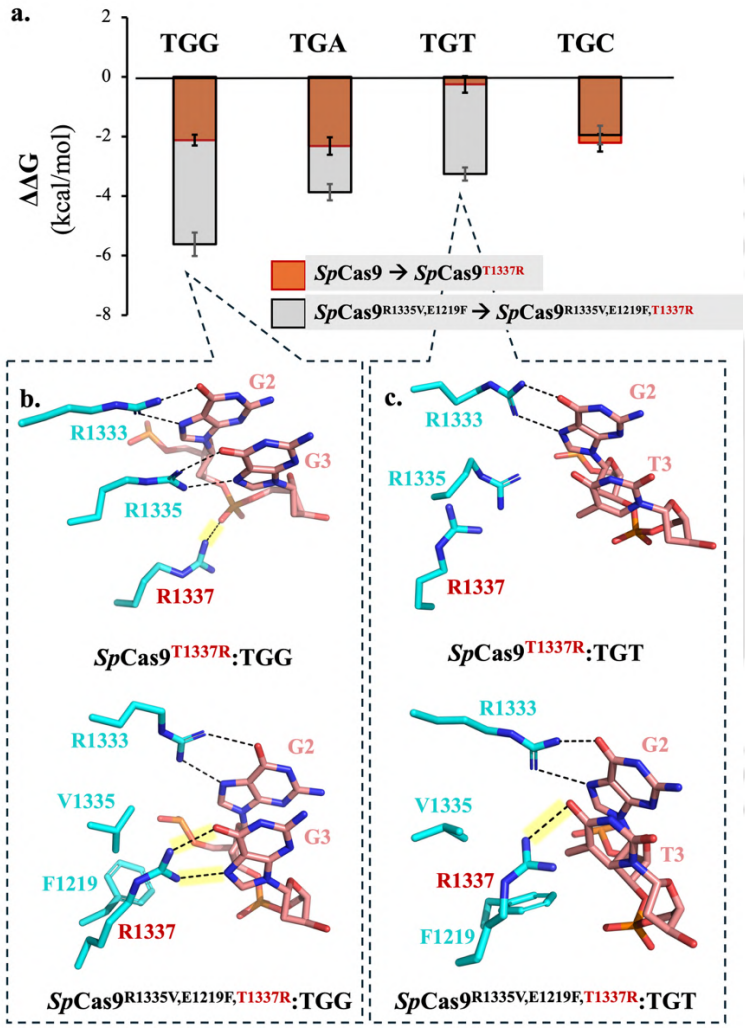


Figure 3.6. (a) Calculated DNA binding free energy difference in response to T1337→R1337 mutation in the wild-type *Sp*Cas9 (orange coloured bar) or double-mutant *Sp*Cas9^{R1335V,E1219F} (grey bar). The dsDNA differs in their PAM sequence (5'-TGG or TGA or TGT or TGC). T1337→R1337 mutation in both *Sp*Cas9 and *Sp*Cas9^{R1335V,E1219F} favour DNA binding, the latter being noticeable (except for 5'-TGC PAM). (b, c)

T1337→R1337 mutation facilitates new protein: DNA interactions (yellow highlighted). The PAM binding pocket of *Sp*Cas9^{R1335V,E1219F} is more hydrophobic than wild-type *Sp*Cas9.

Positively charged amino-acid substitution (G1218→R1218 or A1322→R1322) in *Sp*Cas9 does not form any new protein: DNA interaction (**Figure 3.7**; **appendix Figure A3.10, A3.11.**). Thus, PAM binding affinity is essentially unaffected by G1218→R1218 or A1322→R1322 substitution, which is attributed to significantly lower PAM discrimination (ΔΔG ~ 0 kcal/mol, **Figure 3.2b**).

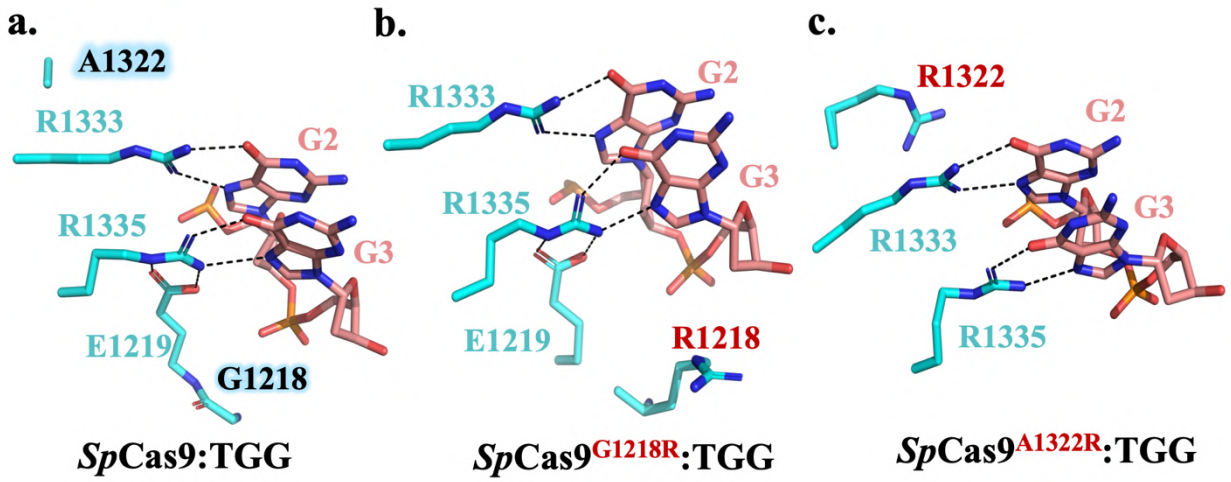


Figure 3.7. Zoomed-in view illustrating (a) MD structure of *Sp*Cas9:TGG, (b) Effect of G1219 → R1218 mutation, (c) Effect of A1322 → R1322 mutation. DNA: protein interaction network (shown in black dotted lines) in the PAM binding pocket is more or less independent of G1218 →R1218 and A1322 → R1322 mutations.

3.4. Discussion

These free energy simulations demonstrate that the location and type of *Sp*Cas9 mutations dramatically influence the energetics of *Sp*Cas9 binding to various DNA targets differing in the PAM sequence (**Figure 3.2b**). As anticipated, the mutation that disrupts the protein-DNA interaction (viz., R1335V) is strongly unfavourable in the *Sp*Cas9:TGG complex. We find that favourable *Sp*Cas9 mutations (E1219F, D1135V, L111R, T1337R) either enhance new protein:

DNA interactions or stabilize existing ones by creating a solvent-excluded low dielectric environment. Experiments (Nishimasu et al., 2018) have confirmed that the R1335A mutant exhibited almost no DNA cleavage activity for the TGG target. However, activity was restored when four surrounding residues (L1111R, T1337R, G1218R, and A1322R) in *Sp*Cas9^{R1335V} were substituted, supporting our findings. The stability of the protein-DNA complex remains unaffected by mutations that do not change the number or strength of the protein-DNA interactions (viz., G1218R, A1322R). These results emphasize that not all seven mutations in Cas9-NG play a critical role in the energetics of protein-DNA binding. Needless to say, the nature of the PAM sequence in the complex is crucial to ensure selectivity. In particular, the R1335V mutation in *Sp*Cas9 strongly disfavours for 5'-TGG or TGA-3' PAM binding, attributing to the loss of protein: DNA interactions (**Figure 3.3**). However, substituting the third purine with a pyrimidine base in the PAM sequence (i.e., TGG or TGA→ TGT or TGC) significantly weakens the selectivity of the complex. Moreover, the protein: PAM interactions (R1333:G2 & R1335:G3) observed in the wild-type *Sp*Cas9:TGG (X-ray as well as in MD structure) are compromised for TGA or TGT or TGC targets (completely lost for TGC target). This also aligns with the experimental observation that *Sp*Cas9 is inactive against TGC PAM (Nishimasu et al., 2018). The high free energy of the *Sp*Cas9:TGC complex drives the dissociation and disrupts catalytic activity.

The increased flexibility of R1335 upon the E1219 → F1219 mutation has been proposed to facilitate new interactions with non-canonical PAM sequences by allowing diverse conformations of the R1335 residue (Guo et al., 2019; Hossain et al., 2025). Interestingly, our observations suggest that the primary role of the E1219F mutation is not to promote new interactions in *Sp*Cas9^{E1219F}: TGG (cognate PAM) but to shield the existing interactions from solvent (**appendix Figure A3.4b**), thereby strengthening the electrostatic interactions and favouring the mutation. The solvent-excluded environment enforced new interactions (**Figure 3.4**, **appendix Table A3.4**) in the *Sp*Cas9^{E1219F}: TGA/TGT/TGC (non-cognate) complexes, particularly boosting the affinity to non-cognate PAM. Non-canonical T-rich PAM readability by a similar mutant *Sp*Cas9^{E1219V} was previously shown to be primarily contributed by the solvent exclusion effect of E1219V mutation (see Chapter 2). E1219V mutation in *Sp*Cas9 was shown to exclude solvents and stabilize the T-rich PAM by providing a hydrophobic cushion to the methyl group of thymine and

promoting new protein: DNA interactions, thereby expanding the PAM readability. The above argument also holds true for the D1335V mutation in *Sp*Cas9, favouring DNA binding (**Figure 3.5 b**; **appendix Figure A3.4c**, **appendix Table A3.5**). It appears that the expansion of the PAM readability of *Sp*Cas9 can be controlled by fine-tuning the hydrophobic environment. Electrostatic interactions (base-specific or non-specific) in the low-dielectric environment improve non-cognate DNA binding, thus expanding PAM readability in *Sp*Cas9. Despite being far away from the PAM binding site ($D1335^{C\beta}$:$G2^{N9}$ = 13 Å, $L1111^{C\beta}$:$G2^{N9}$ = 18.2 Å in PDB 5F9R), D1335V and L111R mutations amplify DNA binding affinity (by ~ 3 kcal/mol) by inducing new non-specific interactions (**Figure 3.5**). Surprisingly, the charge mutation (viz., G1218R or A1322R), which introduces a positive charge in the *Sp*Cas9, does not influence the binding to negatively charged DNA (**Figure 3.7**, **appendix Figure A3.10, A3.11**). Clearly, the mutations can only alter the energetics if they influence protein: DNA interactions. Thus, the location of the mutation is crucial. We propose that G1218R and A1322R mutations in *Sp*Cas9 may not be relevant for expanding PAM readability, thus encouraging experimental verification.

Cas9-NG (a variant of *Sp*Cas9 with seven mutations) is an excellent addition to the CRISPR-Cas9 genome editing toolkit. It has an expanded targeting range, recognizing the NG PAM instead of the traditional NGG while maintaining a similar cleavage potency to the wild-type *Sp*Cas9 (Nishimasu et al., 2018). However, the impact of these mutations on binding affinity was not quantified. It is reasonable to assume that binding affinity is likely directly related to cleavage activity, but this relationship has not been established quantitatively. In experimental studies, DNA cleavage activities indicated *Sp*Cas9-NG can cleave non-canonical targets 5'-TGA, TGT, and TGC-3' more slowly yet effectively as in *Sp*Cas9 (Nishimasu et al., 2018).

In another approach we performed extensive calculations to assess the combined effects of all seven mutations simultaneously (*Sp*Cas9→ Cas9-NG) to include the many-body effect and estimated the ΔΔG (**appendix Figure A3.12, 3.13**). We showed that seven-simultaneous-alchemical transformations in *Sp*Cas9 could reproduce the interaction network of the X-ray structure of Cas9-NG (**appendix Figure A3.13**). Non-canonical PAM (TGA, TGT, TGC) binding is strongly favoured by Cas9-NG, with magnitudes correlating with the experimentally reported change in cleavage activity (Nishimasu et al., 2018). The calculated ΔΔG was then compared with

the "sum of single alchemical transformations" (**appendix Figure A3.12**). Clearly, the magnitude of the ΔΔG from seven simultaneous mutations is 2-5 kcal/mol larger than the sum of single mutations (**appendix Figure A3.12**, indicating potential cooperative effects), but the trends (i.e., the strength of ΔΔG as a function of PAM sequence) remain the same (**Figure 3.8**). Note that a strong correlation between the calculated ΔΔG's and experimental relative cleavage activity is evident (**Figure 3.8**) irrespective of the adopted methodology (simultaneous mutations or sum of single alchemical transformation). The stabilization of the precatalytic complex (i.e., ΔΔG < 0) enhances the cleavage activity of DNA containing non-cognate PAMs, and the strength of ΔΔG is linked to the degree of change in cleavage activity.

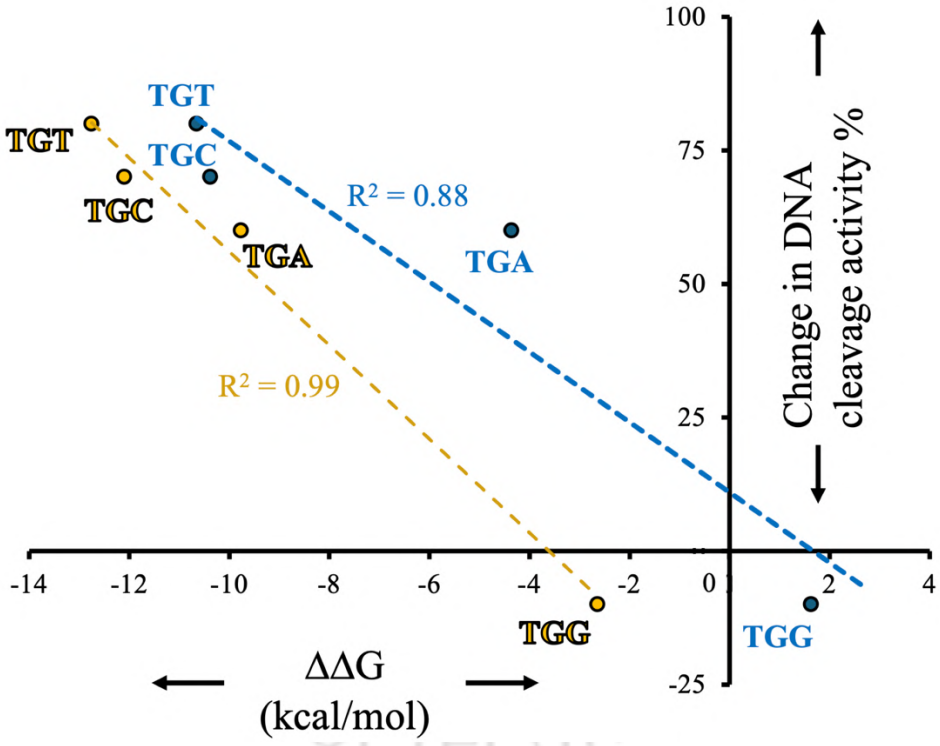


Figure 3.8. Calculated DNA binding free energy difference ($\Delta\Delta G = \Delta G_{bind}^{SpCas9\text{-}NG} - \Delta G_{bind}^{SpCas9}$) versus the difference in DNA cleavage activity (*Sp*Cas9-NG - *Sp*Cas9 in %, obtained from previously documented experiments (Nishimasu et al., 2018)) for various PAM targets (TGG, TGA, TGC, and TGT). ΔΔG was determined by transforming seven residues simultaneously (orange) and compared to the data obtained by summing the energetic contributions from the transformations of individual amino acids (blue). The fitted lines (dotted) for the two data sets indicate a strong correlation.

The possibility of modifying the free energy barrier for the cleavage step in the noncognate complex cannot be dismissed. However, the impact of long-range signalling between the PAM recognition site and distant protease domains, such as HNH and RuvC, which are more than 20 Å apart, appears unnecessary to account for PAM selectivity by *Sp*Cas9. The stability of the pre-catalytic complex appears sufficient to explain PAM decoding. Stabilization (or destabilization) of the non-cognate complex extends (or reduces) PAM readability by increasing (or decreasing) the Boltzmann population.

3.5. Conclusion

Mutations in *Sp*Cas9 can stabilize the pre-catalytic complex (*Sp*Cas9:sgRNA: DNA) through two primary mechanisms: (1) by establishing new non-base-specific interactions between the protein and nucleotides and (2) by reinforcing the electrostatic interactions within a relatively dry and hydrophobic pocket. Stabilization of the complex containing non-cognate PAM sequences allows for broader PAM readability, while destabilization conversely increases the stringency of PAM recognition. The specific positioning of mutations within *Sp*Cas9 is crucial for modulating the stability of this complex. The associated thermodynamic changes (ΔΔG) provide insights into the fidelity of PAM decoding by *Sp*Cas9, thus serving as a foundation for experimental assays assessing the cleavage activity of Cas9.

Chapter 4

Thermodynamics of PAM Selectivity by *Sp*Cas9

This chapter is published in J. Chem. Inf. Model, 2025, 65, 24, 13328–13337

Canonical 5'-NGG-3' PAM recognition in *Sp*Cas9 is often discussed as compatible or incompatible, but its underlying free-energy landscape was never quantified. In this chapter, we evaluated the PAM recognition strength of *Sp*Cas9 by using alchemical free energy calculations, revealing the energetics that influence genome editing accuracy. *Sp*Cas9 does not discriminate at the first position of the NGG sequence, but it penalizes mutations in the second and third positions. *Sp*Cas9 imposes a higher penalty for guanine mutation in the third PAM position compared to the second due to the greater conformational rigidity of R1335 in relation to R1333. Conformational rigidity of R1335 prevents side-chain readjustment for new protein-DNA interactions in non-canonical PAMs. A guanine-to-cytosine substitution in either the second or third position of canonical PAM disrupts direct protein-PAM interactions and leads to solvent exposure. This happens due to strong electrostatic repulsion between the arginine dyad's guanidinium groups and the amine group of cytosine. Interestingly, the strength of *Sp*cas9 in disfavoring a single cytosine substitution (by > 10 kcal/mol) is comparable to that of disfavoring double base substitutions in the NGG sequence. The ability of *Sp*Cas9 to differentiate between non-canonical and canonical PAMs (ΔΔG) is directly related to the number of direct interactions between *Sp*Cas9 and the PAM sequence, as well as the degree of solvent exposure. Loss of direct interactions and increased solvent exposure enhance ΔΔG. The calculated ΔΔG adequately explains the observed differences in DNA cleavage activity of *Sp*Cas9 across various DNA substrates with different PAM sequences. This study connects thermodynamics, structures, and activity to elucidate PAM selectivity in *Sp*Cas9 and may also apply to other CRISPR/Cas systems, offering valuable insights for the rational design of Cas9 variants with modified PAM specificities.

4.1. Background

CRISPR/Cas9 from *Streptococcus pyogenes* Cas9 (*Sp*Cas9) has transformed molecular biology by enabling precise, programmable DNA editing in living cells (Adli, 2018; Jinek et al., 2012). It requires a single guide RNA and a protospacer adjacent motif (PAM), typically 5'-NGG-3', to initiate DNA binding and cleavage (Anders et al., 2014; Nishimasu et al., 2014). PAM recognition triggers DNA unwinding and R-loop formation, enabling double-strand breaks (Anders et al., 2014; Jinek et al., 2014; D. Singh et al., 2016). While strict PAM specificity ensures high fidelity, it limits the range of targetable genomic sites (Guo et al., 2019; Hsu et al., 2014; Kleinstiver et al., 2015).

Efforts to engineer *Sp*Cas9 variants with relaxed or altered PAM readability have largely relied on mutagenesis or directed evolution, yielding variants such as VQR, EQR, VRER (Kleinstiver et al., 2015), xCas9 (Hu et al., 2018), Cas9-NG (Nishimasu et al., 2018), SpG, and SpRY (Walton et al., 2020). These approaches are typically guided by empirical screening (Nishimasu et al., 2018; Walton et al., 2020) rather than a quantitative understanding of the energetic landscape governing PAM recognition. This emphasizes the urgent need to gain mechanistic insights into how PAM mutations influence *Sp*Cas9 binding and activity, which would facilitate the rational design of next-generation genome editors with improved specificity and versatility.

The specificity of PAM recognition in *Sp*Cas9 relies on hydrogen bonding between PAM's guanine bases (G2 and G3 of NGG) and two arginine residues (R1333, R1335 of the PAM-interacting domain (PID, **Figure 4.1a**) (Anders et al., 2014; Nishimasu et al., 2014). X-ray structures (Jiang et al., 2016; Nishimasu et al., 2014) although provide valuable insights into the *Sp*Cas9:PAM interaction but offer limited information on *Sp*Cas9:non-canonical PAM complexes and the energetics of PAM discrimination. High-throughput experiments by Walton et al (Walton et al., 2020), demonstrated that the *Sp*Cas9 editing efficiency drastically reduces in non-canonical PAM sequences, with cytosine substitutions or double base substitutions being the most detrimental. In contrast, adenine substitutions are comparatively tolerated, as reflected by the moderate activity of NAG PAMs and the low activity observed for NGA PAMs. However, the

thermodynamic penalties associated with non-canonical PAMs and how these penalties are structurally encoded remain poorly understood.

This chapter employs structure-based molecular dynamics free energy simulations to quantitatively calculate the energetics of PAM selectivity by *Sp*Cas9 (the canonical 5'-TGG-3' PAM to a diverse set of non-canonical sequences) and highlights the link between energetics, structures, and activity. This integrative approach not only advances our fundamental understanding of the mechanism of CRISPR/Cas9 editing but also highlights the importance of flexibility and solvation for the rational design of genome editing technologies.

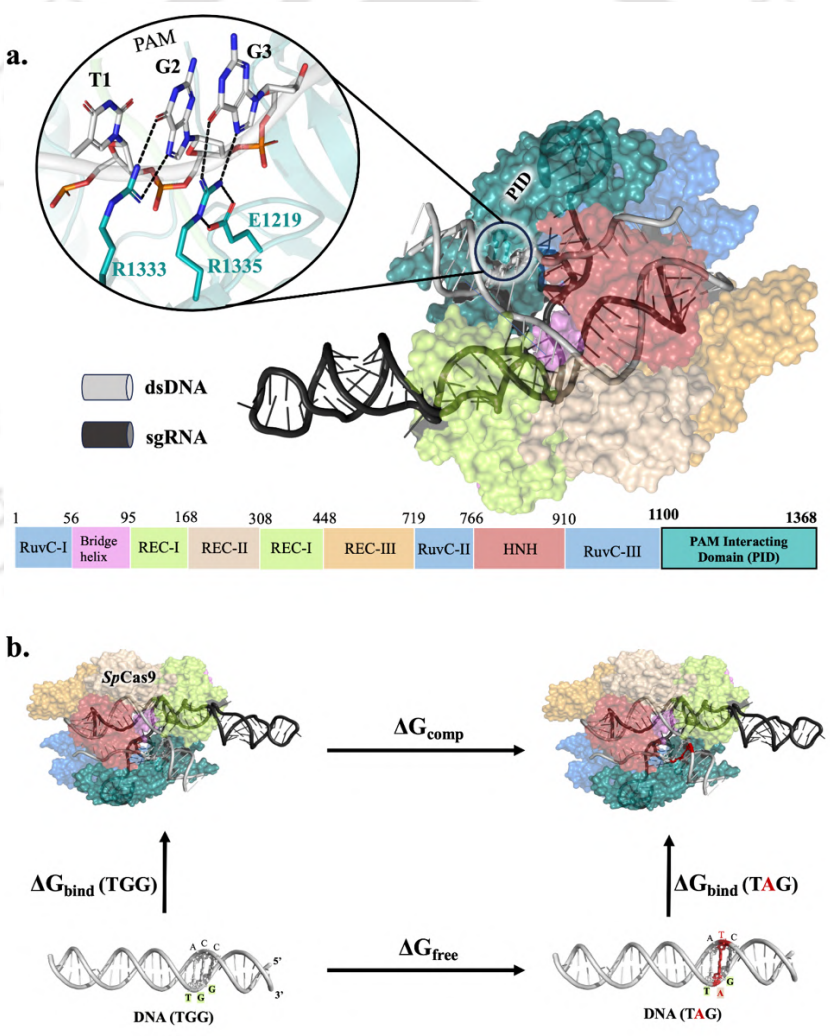


Figure 4.1. (a) Precatalytic *Sp*Cas9 in complex with RNA (black) and target dsDNA (grey)(Jiang et al., 2016). Protein domains are colored as shown in the domain architecture bar below, while nucleic acids are

represented as ribbons. The close-up view of *Sp*Cas9:TGG PAM interactions is illustrated in the circle. Key residues are numbered and represented in stick form, while the interactions are shown with black dotted lines. (b) Thermodynamic cycle for calculating the relative binding free energy (ΔΔG) between different PAM sequences (TGG and non-canonical TAG) to *Sp*Cas9 in the precatalytic complex. Vertical arrows show dsDNA binding to *Sp*Cas9, while horizontal arrows illustrate the alchemical transformation from canonical to non-canonical PAM with *Sp*Cas9 (upper) and in water (lower). Alchemical simulations or base-pair transformation calculated ΔG_{comp} and ΔG_{free} (horizontal legs) and estimated relative binding affinity $\Delta\Delta G = \Delta G_{comp} - \Delta G_{free} = \Delta G_{bind}$ (TAG) - ΔG_{bind} (TGG).

4.2. Methodology

4.2.1. MD setup

The X-ray structure of the precatalytic *Sp*Cas9:sgRNA:dsDNA complex (containing canonical 5'-TGG-3' PAM sequence; PDB 5F9R (Jiang et al., 2016), resolution = 3.4 Å) was retrieved from the Protein Data Bank. The structure of free double-stranded DNA (dsDNA) was modeled using the sequence extracted from the precatalytic *Sp*Cas9 complex (PDB 5F9R). These structures (*Sp*Cas9 precatalytic complex and free dsDNA) were prepared by assigning appropriate topology parameters and adding hydrogen atoms based on physiological pH, generating coordinate files for molecular dynamics simulations. Biomolecules were described using the standard CHARMM36 force field (Huang & Mackerell, 2013; MacKerell et al., 1998), while the TIP3P model (Jorgensen et al., 1998) was used to describe the water molecules. The biomolecule was then positioned at the center and overlaid with an explicit water box with a minimum distance of 1.2 nm (approx. three water layers) between the biomolecule's surface and the box edge to ensure adequate solvation. To neutralize the net charge of the system, monovalent counterions (Na^+) were introduced into the solvated simulation box, which also stabilise the DNA backbone through electrostatic screening.

The details of the box size, solvation, and number of counter ions in the systems are provided in **appendix Table A4.1**. The resulting system was subjected to energy minimization (step size ~0.01 nm, steps ~50000) by employing the steepest descent algorithm (Meza, 2010) until the

maximum force on the system was below 500 kJ/mol/nm, ensuring the removal of steric clashes (if any) and optimizing the initial geometry. The solvated and energy-minimized system was assigned initial velocities generated from a Maxwell–Boltzmann distribution, which was then gradually heated to 300 K over the course of the equilibration phase. The system was then subjected to sequential equilibration comprising in the NVT and NPT ensemble following the protocol elaborated in Chapter 1 (section 1.8.2). Following equilibration, unrestrained production dynamics were performed at 300 K under constant pressure conditions of 1 bar. Temperature was controlled employing velocity rescaling (Bussi et al., 2007) applied to non-hydrogen atoms, with a coupling time constant of 0.1 ps, while pressure was maintained using the Parrinello–Rahman barostat (Nosé & Klein, 1983; Parrinello & Rahman, 1981) with a time constant (τ) of 2 ps. Periodic boundary conditions were applied in all directions throughout the simulations, using a 2 fs integration time step. Van der Waals interactions were truncated at a cutoff distance of 12 Å, while long-range electrostatics were computed using the particle mesh Ewald (PME) method (Darden et al., 1993) with a cutoff of 1.2 nm for short-range electrostatic interactions. The LINCS algorithm was employed to restrain the bond lengths of hydrogen atoms connected to heavy atoms. All simulations were performed using the Gromacs 2023 software (Spoel et al., 2005). Trajectories were saved at 10 ps intervals during the production phase and subsequently used for structural analysis. To enhance sampling, two independent simulation replicas were conducted with distinct initial velocity distributions, yielding a total of 2 μs of trajectory data for analysis. The end-point conformations of the production trajectories served as input for alchemical transformations to estimate relative binding free energies. (as described in the next section).

4.2.2. Alchemical Simulation and Relative Binding Affinity

Alchemical transformations of individual bases within the canonical PAM sequence (5′-TGG-3′) to non-canonical variants (e.g., TGG → TAG), along with their complementary bases, were performed using final structures from conventional MD simulations of both the *Sp*Cas9 precatalytic complex and free double-stranded DNA. An appropriate thermodynamic cycle (**Figure 4.1b**) was employed to compute the change in DNA binding free energy (ΔΔG) resulting

from these base substitutions. In this cycle, the vertical arms represent DNA binding, while the horizontal arms correspond to the alchemical transformation of the PAM sequence, carried out in both the bound state (*Sp*Cas9:DNA complex, upper arrow, **Figure 4.1b**) and the unbound state (free DNA, lower arrow, **Figure 4.1b**). The free energy changes associated with these horizontal arms (ΔG_{comp} and ΔG_{free}) were calculated using the Bennett Acceptance Ratio (BAR) (Bennett, 1976) approach, where the wildtype nucleic acids are slowly transformed to the mutant nucleic acids by tuning the force field parameters. The hybrid structure and topology files of the perturbed nucleic acid residues were generated using the PMX (version 1.2.2) package (Gapsys et al., 2015). Bonds, angles, and dihedrals of the alchemical region were defined using the standard dual-topology scheme of PMX (Gapsys et al., 2015). The standard soft-core potential and λ-scaling in PMX prevented large translations or instabilities during the transformations. A coupling parameter $\lambda \in [0, 1]$ was introduced to smoothly transition between the end states, with $\lambda = 0$ representing the canonical PAM and $\lambda = 1$ representing the non-canonical variant. Intermediate λ values ($0 < \lambda < 1$) correspond to nonphysical hybrid states, which are mixtures of canonical and non-canonical bases (characteristic of alchemical transformations), while connecting two physically releasable states ($\lambda=0$ and $\lambda=1$). These transitions were governed by altering the energy function (U_h, equation 1), representing the hybrid Hamiltonian.

$$U_h = \lambda U_{\text{non-canonical PAM}} + (1- \lambda) U_{\text{canonical PAM}} \quad (1)$$

To ensure smooth transitions between end states, 51 uniformly spaced λ values were used, along with a soft-core potential (Beutler et al., 1994). Alchemical transformations employed the standard GROMACS soft-core potential (Beutler et al., 1994) for both Lennard–Jones and Coulombic terms (sc-alpha = 1.0, sc-power = 1, sc-sigma = 0.3, sc-coul = yes), to ensure smooth nonbonded transitions and reproducible free-energy results. Simulations at each λ point were run for a minimum of 3 ns to a maximum of 10 ns (**appendix Table A4.2**), with the first 1 ns discarded as equilibration to allow the system to adapt to the altered Hamiltonian (Pohorille et al., 2010). Different sampling strategies, ranging from a minimum of 3 ns to a maximum of 10 ns per λ window, were used to ensure the convergence of the free energies (**appendix Table A4.2**). Free energy differences (ΔG_{ij}) between neighbouring λ windows ($\lambda = i$ and $\lambda = j$) were computed via the BAR method (Bennett, 1976) and summed across all windows to obtain the total free energy

change over a horizontal arm. The computations were carried out in both the *Sp*Cas9 precatalytic complex (ΔG_{comp}, upper horizontal arm, **Figure 4.1b**) and the dsDNA free in water (ΔG_{free}, lower horizontal arm, **Figure 4.1b**). The statistical errors related to computed ΔG_{comp} and ΔG_{free} were estimated with the help of the BAR estimator (Bennett, 1976) in Gromacs. The relative binding free energy (ΔΔG) or the binding affinity difference between canonical and non-canonical PAM nucleotides to *Sp*Cas9 was calculated as: $\Delta\Delta G = \Delta G_{comp} - \Delta G_{free}$, and the associated errors for ΔΔG were obtained through standard error propagation of ΔG_{comp} and ΔG_{free}. The sign (positive/negative) of ΔΔG denotes that the mutant (non-canonical PAM) is disfavoured/favoured by *Sp*Cas9, while the magnitude of ΔΔG represents the strength of preference. Good convergence was observed across 3-4 independent replicas, with ΔG values differing by less than 1 kcal/mol and statistical errors remaining below 0.5 kcal/mol (**appendix Table A4.2**), with good overlap of the probability distribution functions between the two neighboring windows (**appendix Figure A4.1**), ensuring reversibility and accuracy of free energy estimates (Pohorille et al., 2010). For transformations involving multiple base changes, simulations were repeated with 101 λ values to validate convergence and robustness, with results comparable to those obtained using 51 λ windows. We confirmed convergence and acceptable statistical uncertainty of the calculated free energies from over 24 μs of molecular dynamics simulation.

4.3. Results

4.3.1. Precatalytic Complex and Free dsDNA in Water

The key *Sp*Cas9:TGG interactions (R1333:G2 and R1335:G3, **Figure 4.1a**) are preserved throughout the MD trajectory (**appendix Figure A4.2a**). The plateau observed in the RMSD versus time plot within 50 ns of MD (**appendix Figure A4.2b**) indicates structural convergence. The low RMSD of approximately 3 Å (**appendix Figure A4.2b**) compared to the X-ray structure confirms that the simulations accurately reproduced the X-ray structure and effectively sampled the relevant conformations. The increased flexibility of the R1333 residue compared to the R1335 residue was evident (**appendix Figure A4.3**) from the MD trajectories of the *Sp*Cas9:TGG complex. Previous studies (Hossain et al., 2025) highlighted that salt-bridge interaction between

R1335 and E1219 (**Figure 4.1a**) lowers the flexibility of R1335 in the canonical *Sp*Cas9:TGG complex, consistent with our observations (**appendix Figure A4.3**). Indeed, we observed high occupancy of R1335:E1219 salt bridge (**appendix Table A4.3**), which resulted in lower flexibility of R1335 compared to R1333 (**appendix Figure A4.3**). Double-stranded DNA (dsDNA) was subjected to conventional and alchemical simulations to study its behaviour in water. The dsDNA was stable during the MD simulations (**appendix Figure A4.2b, d**). Bidentate hydrogen bonds between arginine's guanidinium group and the O6 and N7 positions of guanine are characteristic of the *Sp*Cas9:TGG complex (Hossain et al., 2023, 2025). Thus, guanine substitution in the PAM sequence with other nucleobases disrupts the bidentate interactions (Hossain et al., 2025). Compromise in *Sp*Cas9 activity occurs when the canonical PAM is replaced with a non-canonical analogue, highlighting the importance of the 5'-NGG-3' PAM for cleavage efficiency (Guo et al., 2019; Nishimasu et al., 2018; Walton et al., 2020). In an earlier study, MM/GBSA and thermodynamic integration were employed to estimate TAG versus TGG selectivity in the wild-type and D1135E-mutant of *Sp*Cas9 (Kang et al., 2022). Nevertheless, the comprehensive analysis of PAM recognition in *Sp*Cas9, incorporating various noncanonical sequences, has yet to be quantified.

4.3.2. Structure-based Energetics of PAM recognition by *Sp*Cas9

We quantitatively estimated the impact of PAM substitution on the stability of the *Sp*Cas9 precatalytic complex by performing alchemical free energy calculations using the structures obtained after the production dynamics. These calculations involve computing the change in *Sp*Cas9 binding affinity (ΔΔG) upon mutations in all three positions of TGG PAM. The simulations were performed for single and double-base pair PAM mutations. Calculated binding affinity differences (**Figure 4.2a**) revealed several remarkable features. First, the first position of PAM exhibits a uniformly low discrimination, less than 1 kcal/mol, whereas mutations in the second and third positions are strongly disfavoured, more than 4 kcal/mol (**Figure 4.2a**). This is consistent with the ability of *Sp*Cas9 to recognise any nucleotide in the first position (NGG, where N can be A, T, G, or C), but not for the PAM mutations in the second and third PAM positions. Second, *Sp*Cas9 displayed the highest discrimination for the single cytosine substitution in the second or third position (G2/C2 or G3/C3) of PAM with a strength (ΔΔG ~ 11 kcal/mol, **Figure**

4.2a) comparable to that of double substitutions (C2C3 or A2A3 or T2T3). Previous studies (Walton et al., 2020) indicated that *Sp*Cas9 exhibited no DNA cleavage activity for non-canonical PAM targets containing a single cytosine or a double mutation in the second and/or third position, consistent with the large calculated ΔΔG. Third, the third position of the PAM is relatively more selective (ΔΔG ~ 9 kcal/mol) compared to the second position (ΔΔG ~ 4.5 to 7 kcal/mol), noticeable for G→A substitutions. The TAG PAM is disfavoured by +4.5 kcal/mol compared to +9.0 kcal/mol for TGA PAM in the *Sp*Cas9 complex. Our calculated ΔΔG of +4.5 kcal/mol (TAG versus TGG) aligns with the previously estimated value of +5 kcal/mol by Kang et al (Kang et al., 2022). Moreover, previous experiments (Walton et al., 2020) have demonstrated that *Sp*Cas9 can cleave double-stranded DNA (dsDNA) containing non-canonical TAG PAM with moderate activity, although significantly lower than its activity on the canonical TGG target. Moreover, *Sp*Cas9 exhibits even lower activity on the TGA PAM, which is consistent with the sign and magnitude of the calculated ΔΔG values. Fourth, the *Sp*Cas9 differentiates the mutation in the second position of the PAM, a relatively stronger discrimination disfavouring TTG (by +7 kcal/mol) relative to TAG PAM (by +4.5 kcal/mol). The loss of direct interactions between proteins and the non-canonical PAM (**Figure 4.2b**) led to the exposure of the PAM binding pocket to water (Figure 2c). In several instances (viz. TAG, TCG, TGA, TGT, TGC, and double-base substitutions), the loss of direct interactions in a water-exposed PAM binding pocket is compensated by water-mediated interactions. A direct correlation exists between the magnitude of relative binding free energies (ΔΔG, **Figure 4.2a**) and the loss of direct protein: PAM interactions (**Figure 4.2b**) and solvent exposure of the PAM binding pocket (**Figure 4.2c**).

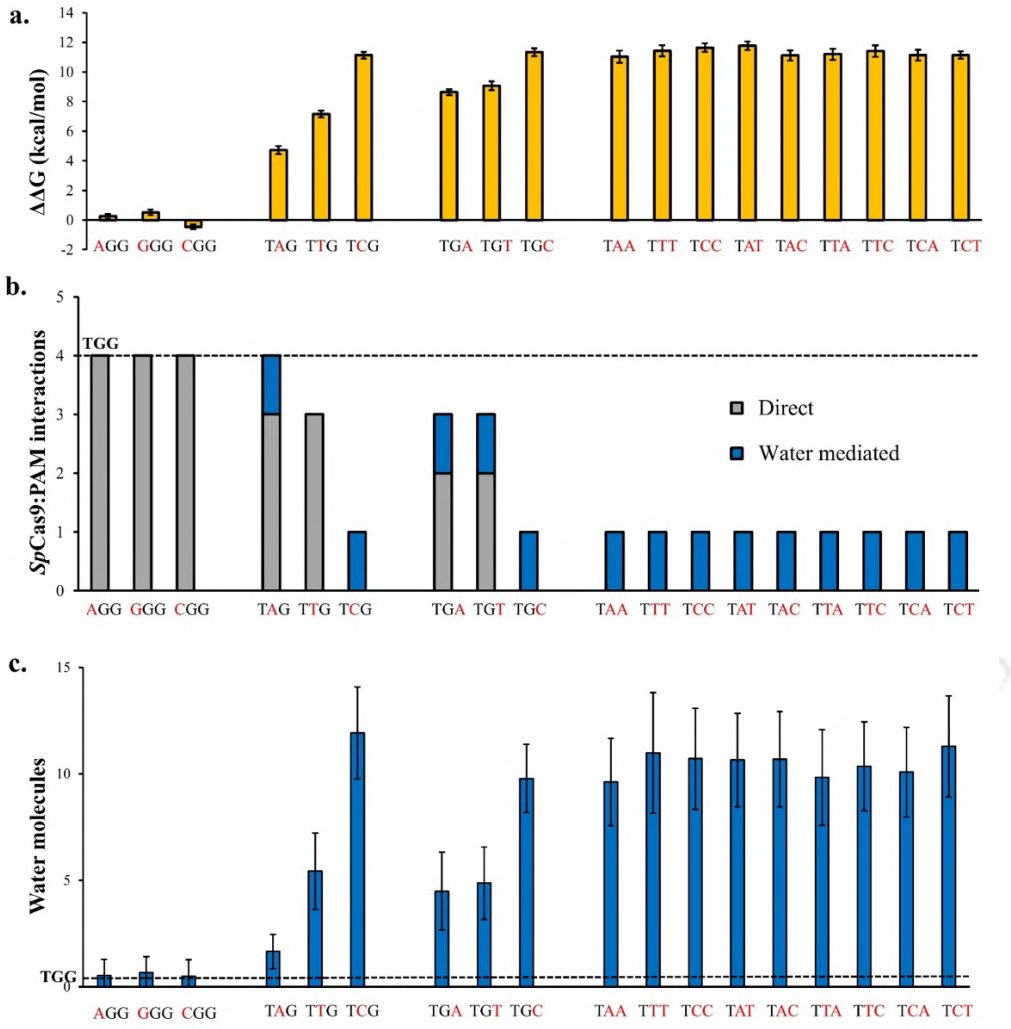


Figure 4.2. Energetics and Structural Determinants of *Sp*Cas9 PAM Recognition. (a) Calculated changes in binding affinity, ΔΔG, for *Sp*Cas9 caused by different base-pair transformations in the canonical PAM sequence (TGG). Error bars represent the standard error of the mean (s.e.m.) calculated across multiple independent simulation trials. (b) Number of direct and water-mediated *Sp*Cas9:non-canonical PAM interactions. Here, the "number of *Sp*Cas9:PAM interactions" refers to the count of interatomic contacts between key heavy atoms of *Sp*Cas9 (NH1, NH2 of R1333 and R1335) and the PAM nucleotides (atoms comprising the Hoogsteen edges) along the MD trajectory. An interaction is considered to occur when the trajectory-averaged interatomic distance between a protein atom and a PAM atom is less than 3.4 Å. (c) Number of water molecules around 3.5 Å of the Hoogsteen edges of the second and third PAM bases, with standard deviation as error. A broken horizontal line indicates interactions and water molecules in the canonical *Sp*Cas9:TGG complex.

4.3.3. Estimated Energetics and Structures

Two arginine residues (R1333 and R1335, **Figure 4.1a**) participate in bidentate interactions with the guanine nucleotides of the canonical *Sp*Cas9:TGG complex. This leads to four direct *Sp*Cas9-PAM interactions. The loss of these direct protein:PAM interactions in the non-canonical PAMs is unfavourable, which is reflected in the positive value of the calculated ΔΔG. The magnitude of ΔΔG is influenced by the number of lost interactions within the non-canonical PAM complexes. It is important to note that the number of direct *Sp*Cas9:PAM interactions is independent of the identity of the first base (**Figure 4.2b**), supporting the classification of NGG as the canonical PAM (where N can be A, T, G, or C). As anticipated, the MD structures show that the first position of the PAM is solvent-exposed and does not have base-specific interactions with *Sp*Cas9 (**Figure 4.3**), further justifying NGG as the canonical PAM.

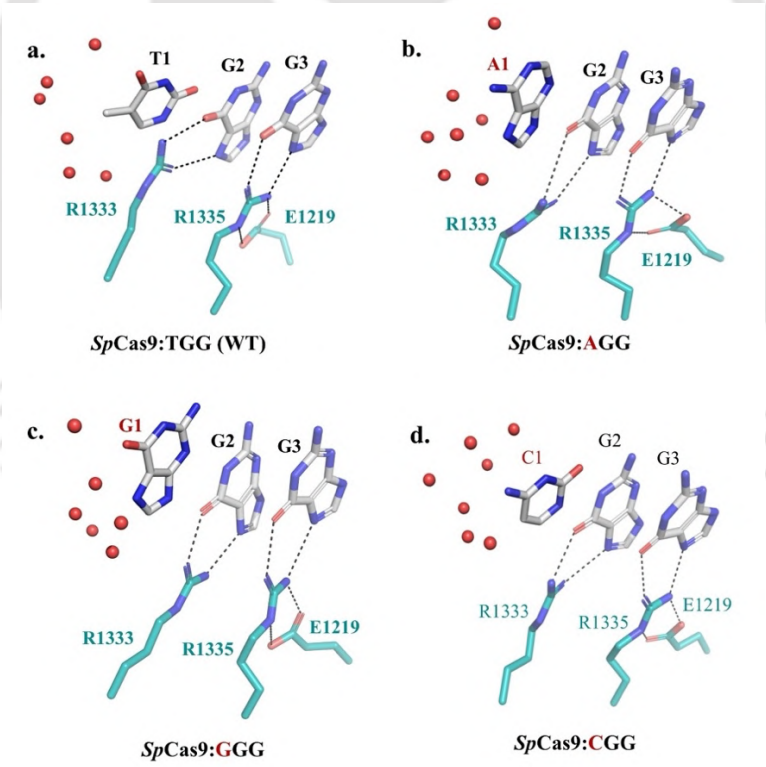


Figure 4.3. MD structures of *Sp*Cas9:PAM complexes (a) TGG, (b) AGG, (c) GGG, (d) CGG. Key residues are represented in sticks, with interactions depicted as dotted lines. Water molecules within 3.5 Å of the Hoogsteen edge of the first base are shown as red spheres. Hydrogen atoms and residue main chains are omitted for clarity.

Substitutions of cytosine in the second or third positions (resulting in TCG or TGC) disrupt all direct interactions, leaving only one water-mediated interaction (**Figures 4.2b and Figure 4.4**, **appendix Table A4.4**). This finding also applies to the double mutant PAMs (TAA, TTT, TCC, TAT, TAC, TTA, TTC, TCA, TCT; **appendix Figure A4.4, appendix Table A4.4**). The electrostatic repulsion between the guanidium group of two arginines (R1333, R1335) and the amine of cytosine disrupted the direct interactions, resulting in the highest solvent exposure of the PAM binding pocket (**Figure 4.2c**).

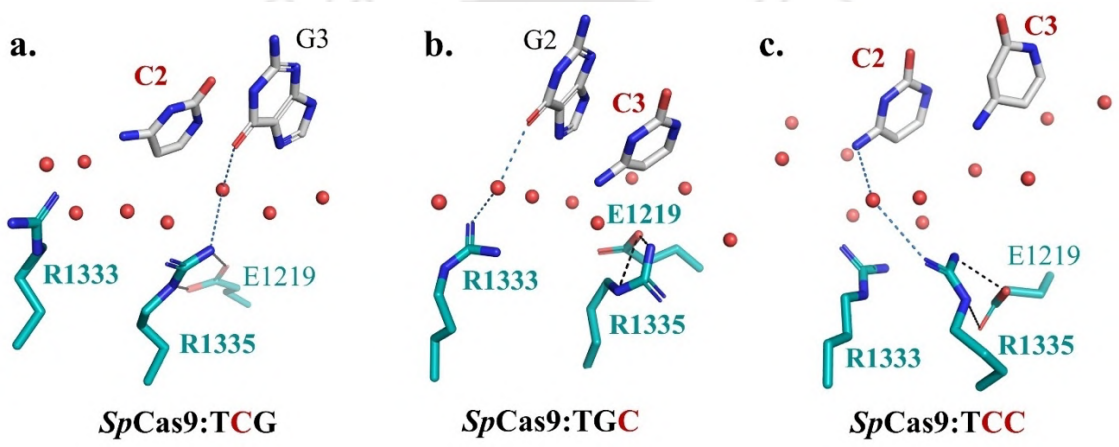


Figure 4.4. Strongest discrimination against cytosine due to the complete loss of direct protein-PAM interactions was observed in the molecular dynamics structures of the precatalytic *Sp*Cas9:non-canonical PAM complexes: (a) TCG, (b) TGC, and (c) TCC. A weak water-mediated interaction between protein and PAM is indicated by a blue dotted line. The interaction between R1335 and E1219 was stable and preserved through the trajectories (represented by a black dotted line). Water molecules within 3.5 Å of the Hoogsteen edge of the second and third bases are shown as red spheres. Hydrogen atoms and residue main chains are omitted for clarity.

Adenine and thymine substitutions in the second (third) position preserved R1335:G3 (R1333:G2) interactions, respectively (**Figure 4.5, appendix Table A4.4**). It is important to note that, despite the mutation at the second position (TAG or TTG, as shown in **Figure 4.5a, b**), one direct interaction between R1333 and DNA remains intact. In contrast, the direct interaction between R1335 and DNA is lost due to the mutation at the third position (TGA or TGT, **Figure 4.5c, d,**

appendix Table A4.4). This difference explains why *Sp*Cas9 exhibits stronger discrimination at the third position. The R1333 was found to be relatively more flexible compared to R1335 (**Figure 4.5, appendix Figure A4.3**). Higher flexibility of R1333 allows the side-chain to bend towards the DNA and interact with the non-canonical bases in the second position (N7 for A2; phosphate for T2). On the other hand, R1335 is rigidified by the E1219, thus unable to adjust its conformation in response to G3/(A3 or T3) mutations, resulting in the loss of direct interaction (TGA and TGT, **Figure 4.5**). This aligns with recent reports (Hossain et al., 2025) that have anticipated the differential flexibility of R1333 and R1335, which is linked to the differing selectivity of the second and third positions, stronger for the latter.

The differential discrimination between TAG and TTG (**Figure 4.2a**) in *Sp*Cas9 warrants further explanation. The key question is why TAG is weakly disfavored (+4.5 kcal/mol) compared to TTG (+7 kcal/mol) by *Sp*Cas9, despite the presence of a salt-bridge interaction for the latter (**Figure 4.5**)? This salt bridge remains stable (**appendix Table A4.4**) throughout the molecular dynamics (MD) trajectories and is a defining feature of the *Sp*Cas9:TTG complex. The strength of the salt-bridge interaction in the *Sp*Cas9:TTG complex is weakened by two main factors. First, we observed that *Sp*Cas9:TTG is relatively more hydrated than *Sp*Cas9:TAG (**Figure 4.2c**). Consequently, the strength of the salt-bridge interaction in *Sp*Cas9:TTG is diminished due to a solvent-exposed high dielectric constant of the local environment. Second, there is a strain in the DNA backbone caused by the salt-bridge interaction (**Figure 4.6**) that was observed from the MD structures. The phosphate backbone of DNA was found to be displaced by ~3.8 Å towards the R1333 to accommodate the new salt-bridge interactions, thus penalising the binding affinity for TTG relative to TAG PAM. Therefore, both the increased solvation of the PAM binding pocket and the induced DNA backbone strain explain why the TTG PAM shows lower affinity than TAG. There is a strong correlation between the solvent exposure of the non-canonical PAM sequences (**Figure 4.2c**) and the calculated relative binding affinities (**Figure 4.2a and 4.2c**). The discriminatory power of *Sp*Cas9 appears to stem from a lack of interactions between the protein and PAM, where the strength or magnitude of ΔΔG can be fine-tuned by controlling the solvent accessibility.

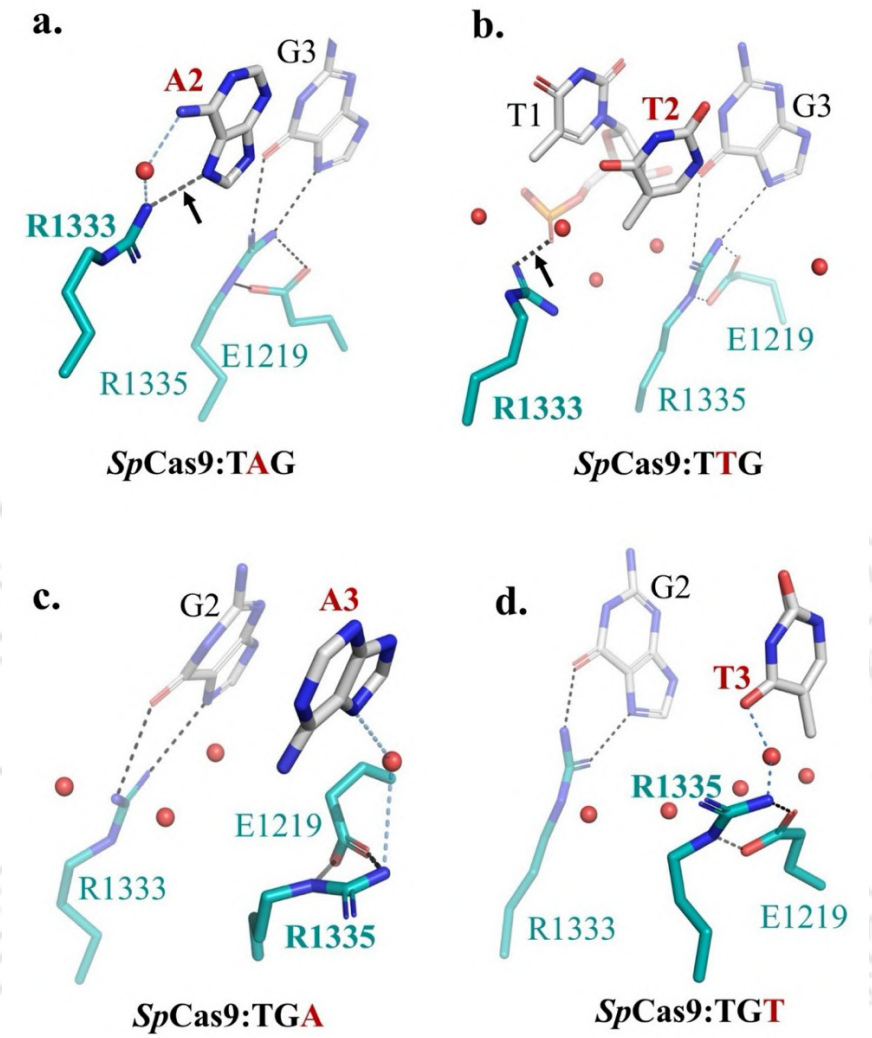


Figure 4.5. Molecular dynamics structures of the non-canonical *Sp*Cas9:PAM complexes: adenine and thymine are present in the second position for (a, b) and in the third position for (c, d). A and T in the second position (top) preserved R1333:DNA direct interactions, highlighted with a black arrow. A new water-mediated interaction (blue dotted lines) in (a) and a characteristic solvent-exposed salt bridge interaction in (b) were observed. The substitution of A and T in the third position does not allow direct interaction with R1335. A relatively larger direct protein: DNA interaction for the second position mutation (top) relative to the third (bottom) is a consequence of the higher flexibility of R1333 relative to R1335. Dotted lines represent interaction. Water molecules within 3.5 Å of the Hoogsteen edge of the second and third bases are shown as red spheres. Hydrogen atoms and residue main chains are omitted for clarity.

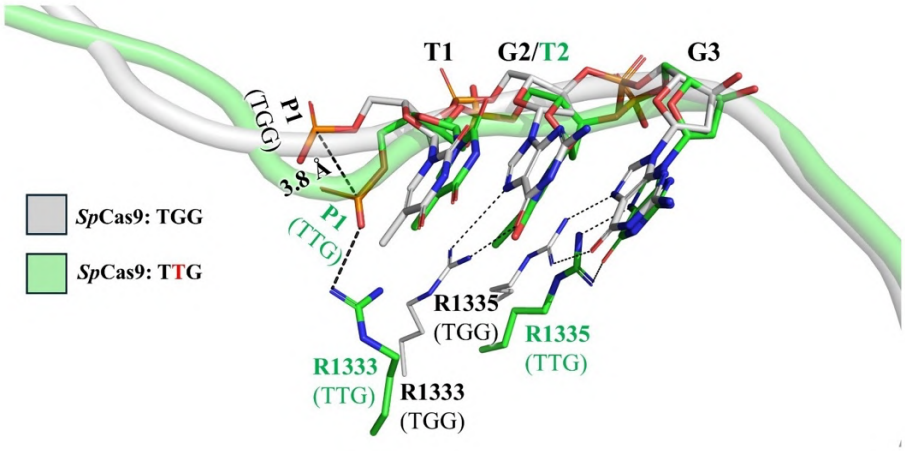


Figure 4.6. Overlay of the DNA backbone of *Sp*Cas9:TGG (grey) and SpCas9:TTG (green) complexes. The key arginine dyad (R1333 and R1335) is highlighted. Salt bridge with R1333 and DNA backbone in *Sp*Cas9:TTG complex caused strain in the DNA backbone, leading to a shift of the DNA phosphate towards the R1333 residue by 3.8 Å relative to the canonical *Sp*Cas9:TGG complex.

4.4. Discussion

We investigated the origin of PAM specificity in *Sp*Cas9 by conducting molecular dynamics free-energy calculations on *Sp*Cas9 with various DNA targets, including both canonical and non-canonical PAM sequences. Experiments demonstrated that PAM recognition followed by R-loop formation (DNA unwinding) in *Sp*Cas9 is the rate-limiting step of DNA cleavage (Hossain et al., 2025; Jones et al., 2017; Nishimasu et al., 2014; Raper et al., 2018; Sternberg et al., 2014). Thus, the sign and magnitude of the calculated PAM preference (ΔΔG) in the precatalytic *Sp*Cas9 complex are crucial for understanding the cleavage activity. We found that *Sp*Cas9 recognises NGG PAMs (where N is A, T, G, or C) with similar affinities (ΔΔG ~ 0 kcal/mol, **Figure 4.2a**), confirming NGG as the canonical PAM (as observed in cleavage assays (Walton et al., 2020)). Both *Sp*Cas9:PAM interactions and solvent accessibility (**Figure 4.2b, 4.2c**) for NGG sequences remain unchanged, resulting in similar affinities. Furthermore, *Sp*Cas9 discriminates strongly (ΔΔG ~ 11 kcal/mol) against cytosine substitution in the second or third position of PAM (**Figure 4.4**), corresponding to a factor of $\sim 10^8$ in terms of affinity. The strong discrimination arises

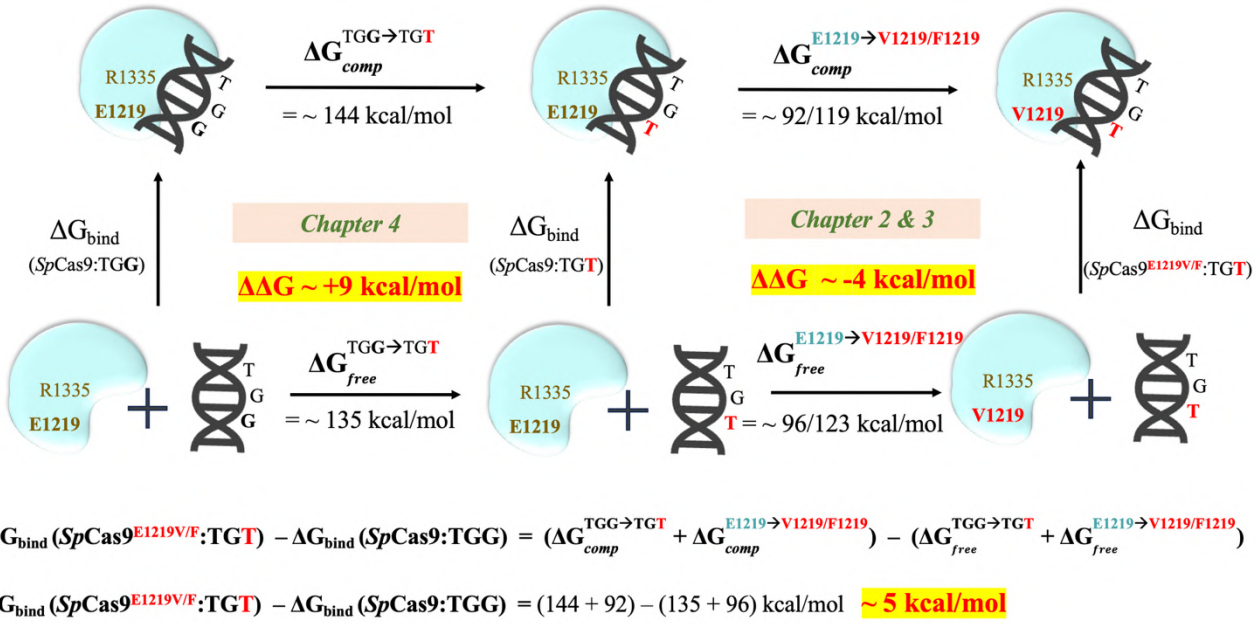
primarily from electrostatic repulsion between the amine group of cytosine and the guanidinium groups of R1333/R1335. Such strong discrimination can dissociate the complex and lower Boltzmann population, which can inhibit the next catalytic step. Loss of protein:PAM interactions and solvent exposure indicate destabilisation of the complex. The high free energy of the non-canonical PAM complexes explains why *Sp*Cas9 was inactive for cytosine-substituted and double-substituted PAM sequences (TAA, TTT, TCC, TAT, TAC, TTA, TTC, TCA and TCT).

Additionally, the rigid conformation of the R1335 residue, which is located near the third PAM position (**Figure 4.5**), prevents nonspecific protein-DNA interactions in the non-cognate PAM complex, thereby enhancing its discriminatory strength, as also anticipated in previous studies (Hossain et al., 2025). The stable interaction between E1219 and R1335 (**appendix Table A4.3**) prevents structural adaptation that would allow bending of R1335 toward the DNA backbone. As a result, the penalty associated with a third-position mutation (TGA and TGT) increases significantly, since no direct interaction occurs in non-canonical pairs (**Figure 4.5**). This enhances the discriminatory power by 2-4 kcal/mol for the third position mutations (TGA and TGT) compared to those in the second position (TAG and TTG). The structural flexibility of R1333 enables interactions with the TAG and TTG non-canonical PAMs, leading to relatively weaker discrimination compared to the third position. Consequently, the discrimination strength of *Sp*Cas9 at the third PAM position is influenced by both electrostatic repulsion and the rigidity of R1335.

The calculated free energy differences show good correlation with the experimental relative activity; a large positive ΔΔG is correlated with poor catalytic activity for non-cognate PAM sequences. The high-throughput experimental editing efficiency assays reported by Walton et al. (Walton et al., 2020), showed that *Sp*Cas9 shows strong cleavage activity for the NGG PAM, moderate activity for NAG, and low activity for NGA. Consistent with these observations, *Sp*Cas9 disfavoured relatively weakly for TAG (by ~4.5 kcal/mol) compared to TGA (~ 7 kcal/mol) with respect to the cognate TGG PAM sequence, explaining its moderate experimental activity for TAG PAM. Similarly, cytosine-containing PAMs and those with double base substitutions showed the highest discrimination (unfavourable), consistent with their experimentally observed inactivity.

Furthermore, the simulations revealed that water accessibility plays a key role in PAM recognition. The accessibility of water in the PAM binding pocket reduces *Sp*Cas9:DNA affinity, a general feature indicated by our results (**Figure 4.2**). The interaction between R1333:N7 (TAG, **Figure 4.5a**) in a dry pocket is stronger than the water-exposed R1333:phosphate salt-bridge interaction (TTG, **Figure 4.5b**), although weaker PAM recognition in the TTG complex is also contributed by local strain in DNA backbone. This causes 2.5 kcal/mol weaker discrimination for TAG compared to TTG when calculated against the canonical TGG PAM sequence. This result clarifies why the cleavage activity of *Sp*Cas9 is higher for a DNA substrate containing TAG than a TTG PAM sequence (Walton et al., 2020). Our results presented in Chapters 2 and 3 as well as some recent studies (Guo et al., 2019; Hossain et al., 2025) showed that *Sp*Cas9 mutations (viz., E1219V, E1219F) expand the non-canonical PAM readability primarily due solvent exclusion of PAM binding pocket and conformational adjustments of flexible R1335. Cleavage assays showed that the *Sp*Cas9^{E1219V} cleaved TGT PAM substrates 60% more effectively than wild-type *Sp*Cas9 (Guo et al., 2019). In Chapters 2 and 3, we reported that the E1219V or E1219F mutation in *Sp*Cas9 favours TGT binding by 4 kcal/mol. Thus, the stabilisation of the TGT substrate in response to the E1219V single mutation in *Sp*Cas9 resulted in the DNA catalysis (**Figure 4.7**). The protein mutation increases the flexibility of R1335 and, most importantly, excludes solvent from the PAM binding cleft. The solvent exclusion effect has two aspects: (1) it stabilises hydrophobic T-rich non-canonical PAM sequences, and (2) it promotes new protein-DNA interactions (either base-specific or non-specific), which reduces the PAM stringency of *Sp*Cas9. The insights from this study urge future engineering efforts to develop CRISPR/Cas9 systems with expanded PAM readability to consider two aspects: (1) increasing the flexibility of R1335, facilitating non-base-specific interactions, and (2) desolvation of the PAM binding cleft for strengthening the protein: DNA interactions. These strategies could significantly enhance the versatility of CRISPR/Cas genome editing, from research to therapy. Our results in Chapter 3 demonstrated that altered electrostatic environments and local hydrophobicity play a crucial role in expanding PAM readability in Cas9-NG variant. These suggest that the mechanistic insights obtained here for *Sp*Cas9 are likely generalizable to rational engineering of PAM recognition across Cas9 and Cas12 orthologs. The simulations effectively illustrate the accuracy of PAM decoding by wild-type *Sp*Cas9, allowing for a thorough examination of the mechanism in terms

of energetics and its link to structures and activity. The thermodynamics underlying the PAM recognition by wild-type *Sp*Cas9 explain its varying DNA cleavage activity on DNA substrates with different PAM (canonical and non-canonical) sequences.



$$\Delta G_{bind}(SpCas9^{E1219V/F}{:}TGT) - \Delta G_{bind}(SpCas9{:}TGG) = (\Delta G_{comp}^{TGG\rightarrow TGT} + \Delta G_{comp}^{E1219\rightarrow V1219/F1219}) - (\Delta G_{free}^{TGG\rightarrow TGT} + \Delta G_{free}^{E1219\rightarrow V1219/F1219})$$

$$\Delta G_{bind}(SpCas9^{E1219V/F}{:}TGT) - \Delta G_{bind}(SpCas9{:}TGG) = (144 + 92) - (135 + 96)\ \text{kcal/mol} \sim 5\ \text{kcal/mol}$$

Figure 4.7. Thermodynamic cycle: *Sp*Cas9:TGG versus *Sp*Cas9^{E1219V}:TGT binding. Based on the findings from this study (appendix Table A4.2) and results of Chapters 2 and 3, the estimated difference in binding affinity between *Sp*Cas9:TGG and *Sp*Cas9^{E1219V}:TGT is +5 kcal/mol, favouring the former. Wild-type *Sp*Cas9 showed moderate DNA catalysis on the TAG substrate (Walton et al., 2020), which is less favourable by +4.5 kcal/mol compared to the canonical TGG substrate (Figure 4.2). Thus, a free energy difference of +5 kcal/mol between *Sp*Cas9^{E1219V}:TGT and *Sp*Cas9:TGG suggests that DNA catalysis for the former is plausible, corroborating to the reported catalysis assay (Guo et al., 2019).

4.5. Conclusion

The strict requirement for an "NGG" PAM sequence for *Sp*Cas9 activity stems from its varying affinity for canonical (NGG) versus non-canonical PAM sequences. The loss of base-specific interactions between the protein and the PAM, along with the solvent exposure of the PAM binding cleft, reduces the likelihood of non-canonical PAM binding, ensuring a high level of

stringency. The rigid R1335 conformation enhances the discriminatory power of *Sp*Cas9. Our findings connect the thermodynamics, structures, and cleavage activity, offering a framework for the rational design of Cas9 variants with tailored specificity. Thus, it can be argued that mutations in Cas9 that facilitate non-base-specific interactions within a desolvated pocket may reduce the PAM stringency in *Sp*Cas9. Targeted site-directed mutagenesis, followed by in vitro cleavage assays, may quantify the impact of specific mutations on PAM selectivity and validate our predictions.

Chapter 5

Overall Conclusion and Future Prospects

In recent years, several engineered variants of Cas9 with relaxed PAM specificities have been developed, including xCas9 (Hu et al., 2018), Cas9-NG (Nishimasu et al., 2018), SpG (Walton et al., 2020), and SpRY (Walton et al., 2020). However, the development of these variants has largely relied on empirical or semi-empirical methods, such as random mutagenesis and directed evolution, without a detailed quantitative understanding of the energetic landscape involved in PAM recognition. The molecular mechanisms and the energetics associated with how mutations in engineered variants expand PAM recognition have not been sufficiently explored. This thesis aims to address this knowledge gap. It quantitatively estimates how specific mutations either in the PAM or in *Sp*Cas9 affect the stability of the *Sp*Cas9-nucleotide complex. It explains how wild-type *Sp*Cas9 distinguishes between canonical PAM sequences (5′-NGG-3′) and non-canonical ones, as well as how mutations in *Sp*Cas9 affect DNA binding (**Figure 5.1**). Alchemical free-energy calculations were employed to quantify the structural, dynamic, and energetic determinants of *Sp*Cas9: PAM recognition. The thesis quantitatively examined (i) the impact of a single mutation E1219V (**Chapter 2**), (ii) the thermodynamic basis of broadened PAM compatibility in the engineered Cas9-NG variant, which contains seven mutations in the PAM-interacting domain (**Chapter 3**), and (iii) the energetic determinants that allow wild-type *Sp*Cas9 to selectively recognize the canonical NGG PAM while rejecting non-cognate PAMs (**Chapter 4**). The results revealed that *Sp*Cas9:PAM selectivity is governed by a combination of local hydrophobicity, electrostatic interactions, and the conformational flexibility of the PAM-binding cleft of *Sp*Cas9. The increased flexibility of *Sp*Cas9 enables the formation of new interactions (mostly non-specific) with DNA containing non-cognate PAM sequences, while increased local hydrophobicity enhances the strength of these interactions (due to the low local dielectric). The study establishes a connection between estimated energetics and molecular structures, and

explains the molecular mechanisms underlying experimentally observed differences in cleavage activity of *Sp*Cas9. These collectively shape the energetic landscape of recognition, offering mechanistic insights crucial for rational Cas9 engineering.

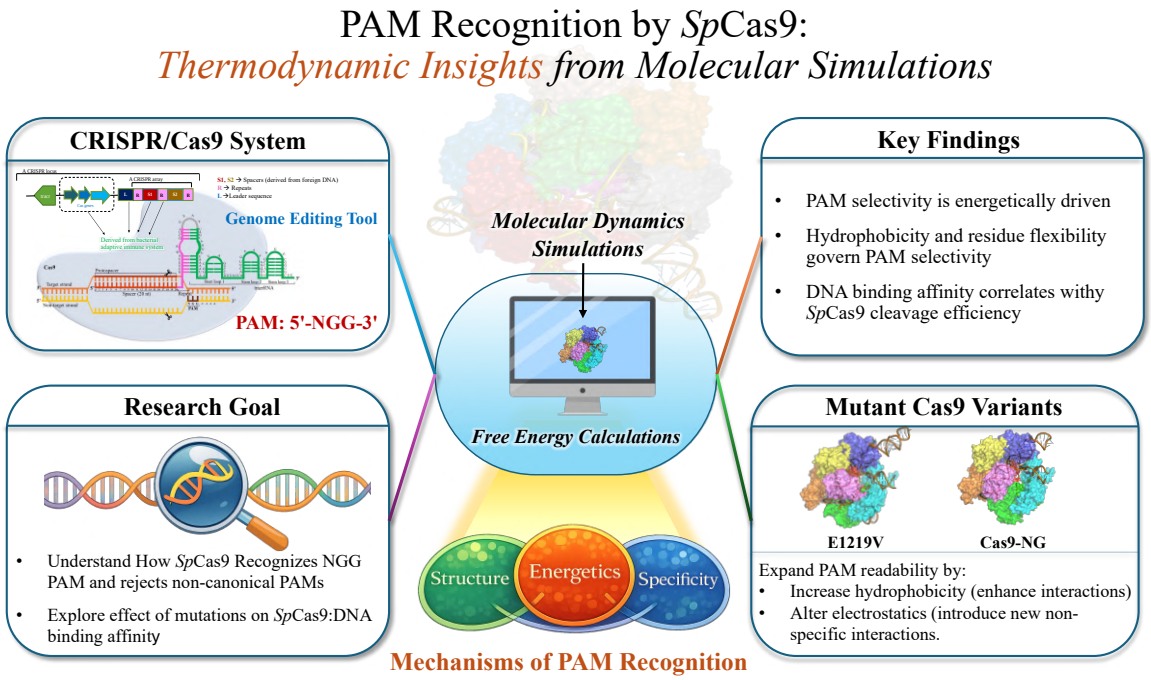


Figure 5.1. Schematic representation demonstrating interplay between structure, energetics, and specificity underlying PAM recognition by *Sp*Cas9. Results from this thesis demonstrate that alternating the local residue flexibility and hydrophobicity can reshape the free-energy landscape of Cas9:DNA interactions, thereby modulating DNA binding affinity and cleavage efficiency.

5.1. Summary of Key Findings

Chapter 1 of the thesis introduced the CRISPR/Cas9 genome editing mechanism, the importance of PAM recognition and formulated the research questions and three key objectives.

The first objective (**Chapter 2**) explored the energetic basis of how a single E1219V mutation broadens PAM readability from the canonical TGG to T-rich PAMs, such as TGT and GAT, and provided a molecular explanation for the experimentally observed broadened PAM compatibility of E1219V. The E1219V mutation in *Sp*Cas9 favoured non-canonical PAM binding by

approximately 2 to 4 kcal·mol^{-1} by promoting non-specific interactions in a hydrophobic environment, thereby reducing PAM stringency. Previously, an experimental study confirmed a complete loss of activity for the R1335A mutation. We hypothesised that a double mutant *Sp*Cas9^{R1335A,E1219V,} may exhibit cleavage activity compared to the inactive R1335A variant, which is subject to future experimental validations.

The second objective (**Chapter 3**) examined the thermodynamic origins of the broadened PAM recognition exhibited by the engineered Cas9-NG variant (*Sp*Cas9$^{R1335V, E1219F, D1135V, L1111R, T1337R, G1218R, A1322R}$), which contains seven mutations within the PAM-interacting domain. The findings indicated that the seven mutations in Cas9-NG are not necessary for relaxed PAM recognition and impact of each mutation on PAM binding is highly dependent on its nature and location in *Sp*Cas9. Five mutations (R1335V, E1219F, D1135V, L1111R, T1337R) may be sufficient. As discussed in Chapter 2, the mutations stabilise the *Sp*Cas9:non-canonical PAM complex, either by introducing new non-base-specific SpCas9-DNA interactions or strengthening the electrostatic interactions in the dry and hydrophobic pocket. The study demonstrated an excellent correlation between estimated relative binding affinity and relative cleavage efficiency, thereby highlighting the DNA binding step as the most crucial in the *Sp*Cas9 editing pathway.

The third objective (**Chapter 4**) discussed the effect of PAM mutation on the *Sp*Cas9 binding and addressed the mechanism by which *Sp*Cas9 discriminates against non-cognate PAMs. We demonstrated that a canonical PAM (5'-TGG-3') mutation at the second and third positions incur penalties that are proportional to the loss of base-specific contacts and an increase in hydration within the PAM binding pocket. This leads to reduced stability of non-canonical PAM complexes, which ensures strict selectivity for PAM sequences. The rigid R1335 conformation enhances the discriminatory power of *Sp*Cas9, leading to third-position substitution being more penalised than substitutions in the second position. Furthermore, cytosine substitution in the PAM is heavily penalised by *Sp*Cas9, resulting in a complete loss of interaction and increased solvent exposure.

The computational results presented in this thesis show excellent agreement with experimentally observed cleavage efficiencies of *Sp*Cas9 and Cas9-NG across different substrates. The associated thermodynamic changes (ΔΔG) provide insights into the fidelity of PAM decoding by *Sp*Cas9,

thus serving as a foundation for experimental assays assessing the cleavage activity of Cas9. Our findings connect the thermodynamics, structures, and cleavage activity, offering a framework for the rational design of Cas9 variants with tailored specificity.

5.2. Take home message

This thesis introduces a thermodynamic framework for understanding PAM recognition in wild-type *Sp*Cas9 and its engineered variants. It reveals that PAM specificity arises from a balance of local residue flexibility and hydrophobicity within the PAM-binding cleft, rather than just hydrogen bonding. The findings clarify how targeted mutations can adjust this balance to either increase PAM stringency or expand PAM recognition. Stabilising pre-catalytic complexes with non-cognate PAMs enhances readability, while destabilising them increases discrimination. It seems that an increase in hydrophobicity, non-specific interactions, and flexibility of the PAM binding pocket of *Sp*Cas9 could potentially expand the PAM readability of *Sp*Cas9. The alchemical free energy calculations employed in this thesis successfully bridge structural, thermodynamic, and experimental observations. This integrative strategy not only accounts for existing data but also generates testable hypotheses for future experimental investigation.

5.3. Scope and Limitations

This thesis establishes structure-based free energy simulations as a practical and reliable approach for addressing key questions about the specificity and accuracy of protein:nucleotide recognitions in biological systems. Nevertheless, the computational framework employed here carries certain inherent methodological limitations, like the use of fixed-charge force fields, simplified treatments of ions and solvent, and the challenges associated with modelling long-range electrostatic effects in truncated simulation models. In addition, the analysis is carried out assuming the system in in thermodynamic equilibrium and does not explicitly model downstream kinetic processes such as R-loop propagation, conformational activation, or DNA cleavage. While the possibility that the free-energy barrier associated with the cleavage step is altered in non-canonical complexes cannot be ruled out, the strong correlation observed between binding affinity

and cleavage activity suggests that PAM recognition largely governs cleavage efficiency, and that the stability of the pre-catalytic complex is sufficient to explain the experimentally observed trends in *Sp*Cas9 activity. Although computational predictions are compared with available experimental data, the simulations do not substitute for direct experimental data, and the hypotheses generated in this thesis warrant future experimental validation. Future improvements could include the use of polarizable force fields or hybrid QM/MM approaches to better capture electronic and long-range effects, as well as investigate how mutations influence the catalytic step of *Sp*Cas9. Taken together, these considerations define the scope of the present work while preserving the validity of the central conclusion that PAM recognition is governed by energetically driven selection mechanisms.

5.4. Future Prospects

Several promising research directions emerge from this work:

1. **Experimental validations**: Targeted site-directed mutagenesis followed by in vitro and cell-based cleavage assays can quantify the impact of specific mutations on PAM selectivity and validate the hypotheses generated in this thesis.
2. **Energetics along the *Sp*Cas9 editing pathway**: Future studies may quantify how mutations affect both the binding step and catalytic activation steps of the *Sp*Cas9 genome editing pathway, thus providing a complete energetic profile of the editing pathway.
3. **Application to Other Cas Orthologs**: The methodology could be extended to other Cas9 homologues like *Sa*Cas9, *Nm*Cas9 or other Cas family proteins like Cas12a/Cas12b and Cas13 to derive generalizable design principles across CRISPR families. The methodology could also be extended to other nucleotide binding proteins like ribosomes, polymerases, transcription factors etc.
4. **Understanding the mechanism of off-target effects in *Sp*Cas9**: This study can be expanded to examine how local flexibility, hydration, and non-specific interactions affect off-target binding and cleavage. By mapping the changes in $\Delta\Delta G$ associated with mismatches at different locations in the target DNA, we can investigate the thermodynamic origins of off-target activity.

References

Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., Verdine, V., Cox, D. B. T., Kellner, M. J., Regev, A., Lander, E. S., Voytas, D. F., Ting, A. Y., & Zhang, F. (2017). RNA targeting with CRISPR–Cas13. *Nature 2017 550:7675*, *550*(7675), 280–284. https://doi.org/10.1038/nature24049

Adane, M., & Alamnie, G. (2024). CRISPR/Cas9 mediated genome editing for crop improvement against Abiotic stresses: current trends and prospects. *Functional & Integrative Genomics*, *24*(6), 199. https://doi.org/10.1007/S10142-024-01480-2

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. *Nature Communications 2018 9:1*, *9*(1), 1–13. https://doi.org/10.1038/s41467-018-04252-2

Ahmed, M., Daoud, G. H., Mohamed, A., & Harati, R. (2021). New Insights into the Therapeutic Applications of CRISPR/Cas9 Genome Editing in Breast Cancer. *Genes 2021, Vol. 12, Page 723*, *12*(5), 723. https://doi.org/10.3390/GENES12050723

Ahsan, M., Nierzwicki, Ł., East, K. W., Binz, J., Hsu, R. V., Arantes, P. R., Skeens, E., Pacesa, M., Jinek, M., Lisi, G. P., & Palermo, G. (2023). Principles of DNA cleavage in CRISPR-Cas9. *Biophysical Journal*, *122*(3), 170a. https://doi.org/10.1016/J.BPJ.2022.11.1068

Allnér, O., & Nilsson, L. (2011). Nucleotide modifications and tRNA anticodon–mRNA codon interactions on the ribosome. *RNA*, *17*(12), 2177. https://doi.org/10.1261/RNA.029231.111

Allnér, O., Nilsson, L., & Villa, A. (2012). Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations. *Journal of Chemical Theory and Computation*, *8*(4), 1493–1502. https://doi.org/10.1021/CT3000734

Almlöf, M., Andér, M., & Åqvist, J. (2007). Energetics of codon-anticodon recognition on the small ribosomal subunit. *Biochemistry*, *46*(1), 200–209. https://doi.org/10.1021/BI061713I

Altermann, E., Russell, W. M., Azcarate-Peril, M. A., Barrangou, R., Buck, B. L., McAuliffe, O., Souther, N., Dobson, A., Duong, T., Callanan, M., Lick, S., Hamrick, A., Cano, R., & Klaenhammer, T. R. (2005). Complete genome sequence of the probiotic lactic acid bacterium Lactobacillus acidophilus NCFM. *Proceedings of the National Academy of Sciences*, *102*(11), 3906–3912. https://doi.org/10.1073/PNAS.0409188102

Amitai, G., & Sorek, R. (2016). CRISPR–Cas adaptation: insights into the mechanism of action. *Nature Reviews Microbiology 2016 14:2*, *14*(2), 67–76. https://doi.org/10.1038/nrmicro.2015.14

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

Andrew McCammon, J. (1991). Free energy from simulations: Current Opinion in Structural Biology 1991, 1 : 196–200. *Current Opinion in Structural Biology*, *1*(2), 196–200. https://doi.org/10.1016/0959-440X(91)90061-W

Arantes, P. R., Mitchell, B. P., Saha, A., Nierzwicki, L., Pacesa, M., Jinek, M., & Palermo, G. (2023). Structure and dynamics of off-target effects in CRISPR-Cas9. *Biophysical Journal*, *122*(3), 190a. https://doi.org/10.1016/J.BPJ.2022.11.1165

Barrangou, R., & Doudna, J. A. (2016). Applications of CRISPR technologies in research and beyond. *Nature Biotechnology 2016 34:9*, *34*(9), 933–941. https://doi.org/10.1038/nbt.3659

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, *315*(5819), 1709–1712. https://doi.org/10.1126/SCIENCE.1138140

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, *22*(2), 245–268. https://doi.org/10.1016/0021-9991(76)90078-4

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R., & van Gunsteren, W. F. (1994). Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, *222*(6), 529–539. https://doi.org/10.1016/0009-2614(94)00397-1

Bhardwaj, A., Tomar, P., & Nain, V. (2024). Machine Learning-Driven Prediction of CRISPR-Cas9 Off-Target Effects and Mechanistic Insights. *Eurobiotech Journal*, *8*(4), 213–229. https://doi.org/10.2478/EBTJ-2024-0020

Bhati, A. P., Wan, S., Wright, D. W., & Coveney, P. V. (2017). Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *Journal of Chemical Theory and Computation*, *13*(1), 210–222. https://doi.org/10.1021/ACS.JCTC.6B00979

Bhushan, B., Singh, K., Kumar, S., & Bhardwaj, A. (2024). Advancements in CRISPR-Based Therapies for Genetic Modulation in Neurodegenerative Disorders. *Current Gene Therapy*, *25*(1), 34–45. https://doi.org/10.2174/0115665232292246240426125504

Bogusz, S., Cheatham, T. E., & Brooks, B. R. (1998). Removal of pressure and free energy artifacts in charged periodic systems via net charge corrections to the Ewald potential. *The Journal of Chemical Physics*, *108*(17), 7070–7084. https://doi.org/10.1063/1.476320

Bolotin, A., Quinquis, B., Sorokin, A., & Dusko Ehrlich, S. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, *151*(8), 2551–2561. https://doi.org/10.1099/MIC.0.28048-0

Braga, L. A. M., Filho, C. G. C., & Mota, F. B. (2022). Future of genetic therapies for rare genetic diseases: what to expect for the next 15 years?: *Https://Doi.Org/10.1177/26330040221100840*, *3*, 263300402211008. https://doi.org/10.1177/26330040221100840

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

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., … Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, *30*(10), 1545–1614. https://doi.org/10.1002/JCC.21287

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, *4*(2), 187–217. https://doi.org/10.1002/JCC.540040211

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V., & Van Der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, *321*(5891), 960–964. https://doi.org/10.1126/SCIENCE.1159689

Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *Journal of Chemical Physics*, *126*(1). https://doi.org/10.1063/1.2408420/186581

Carroll, D. (2012). A CRISPR Approach to Gene Targeting. *Molecular Therapy*, *20*(9), 1658–1660. https://doi.org/10.1038/MT.2012.171

Casalino, L., Nierzwicki, Ł., Jinek, M., & Palermo, G. (2020). Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR-Cas9 Revealed by Ab Initio Molecular Dynamics. *ACS Catalysis*, *10*(22), 13596–13605. https://doi.org/10.1021/ACSCATAL.0C03566

Cheatham, T. E., & Case, D. A. (2013). Twenty-five years of nucleic acid simulations. *Biopolymers*, *99*(12), 969–977. https://doi.org/10.1002/BIP.22331

Chen, H., Maia, J. D. C., Radak, B. K., Hardy, D. J., Cai, W., Chipot, C., & Tajkhorshid, E. (2020). Boosting free-energy perturbation calculations with GPU-Accelerated NAMD. *Journal of Chemical Information and Modeling*, *60*(11), 5301–5307. https://doi.org/10.1021/ACS.JCIM.0C00745

Chen, L., Brügger, K., Skovgaard, M., Redder, P., She, Q., Torarinsson, E., Greve, B., Awayez, M., Zibat, A., Klenk, H. P., & Garrett, R. A. (2005). The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota. *Journal of Bacteriology*, *187*(14), 4992–4999. https://doi.org/10.1128/JB.187.14.4992-4999.2005

Chipot, C., & Pohorille, A. (2007). Calculating free energy differences using perturbation theory. *Springer Series in Chemical Physics*, *86*, 33–75. https://doi.org/10.1007/978-3-540-38448-9_2

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science*, *339*(6121), 819–823. https://doi.org/10.1126/SCIENCE.1231143

Cournia, Z., & Chipot, C. (2024). Applications of Free-Energy Calculations to Biomolecular Processes. A Collection. *The Journal of Physical Chemistry B*, *128*(14), 3299–3301. https://doi.org/10.1021/ACS.JPCB.4C01283

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, *46*(W1), W246–W251. https://doi.org/10.1093/NAR/GKY425

Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A., & Yildiz, A. (2017). A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Science Advances*, *3*(8). https://doi.org/10.1126/SCIADV.AAO0027

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *Citation: J. Chem. Phys*, *98*, 10089. https://doi.org/10.1063/1.464397

Das, A., Rai, J., Roth, M. O., Shu, Y., Medina, M. L., Barakat, M. R., & Li, H. (2023). Coupled catalytic states and the role of metal coordination in Cas9. *Nature Catalysis 2023 6:10*, *6*(10), 969–977. https://doi.org/10.1038/s41929-023-01031-1

Deb, S., Choudhury, A., Kharbyngar, B., & Satyawada, R. R. (2022). Applications of CRISPR/Cas9 technology for modification of the plant genome. *Genetica*, *1*, 1–12. https://doi.org/10.1007/S10709-021-00146-2

Deift, P., & Zhou, X. (1993). A Steepest Descent Method for Oscillatory Riemann--Hilbert Problems. Asymptotics for the MKdV Equation. *The Annals of Mathematics*, *137*(2), 295. https://doi.org/10.2307/2946540

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature 2011 471:7340*, *471*(7340), 602–607. https://doi.org/10.1038/nature09886

Estarellas, C., Otyepka, M., Koča, J., Banáš, P., Krepl, M., & Šponer, J. (2015). Molecular dynamic simulations of protein/RNA complexes: CRISPR/Csy4 endoribonuclease. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1850*(5), 1072–1090. https://doi.org/10.1016/J.BBAGEN.2014.10.021

Feller, S. E., Zhang, Y., Pastor, R. W., & Brooks, B. R. (1995). Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics*, *103*(11), 4613–4621. https://doi.org/10.1063/1.470648

Fizikova, A., Tikhonova, N., Ukhatova, Y., Ivanov, R., Khlestkina, E., Amjad Nawaz, M., Golokhvast, K. S., Chung, G., Tsatsakis, A. M., & Antoniou, M. N. (2021). Applications of CRISPR/Cas9 System in Vegetatively Propagated Fruit and Berry Crops. *Agronomy 2021, Vol. 11, Page 1849*, *11*(9), 1849. https://doi.org/10.3390/AGRONOMY11091849

Frangoul, H., Altshuler, D., Cappellini, M. D., Chen, Y.-S., Domm, J., Eustace, B. K., Foell, J., de la Fuente, J., Grupp, S., Handgretinger, R., Ho, T. W., Kattamis, A., Kernytsky, A., Lekstrom-Himes, J., Li, A. M., Locatelli, F., Mapara, M. Y., de Montalembert, M., Rondelli, D., … Corbacioglu, S. (2021). CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *The New England Journal of Medicine*, *384*(3), 252–260. https://doi.org/10.1056/NEJMOA2031054

Frenkel, Daan., & Smit, Berend. (2023). *Understanding molecular simulation : from algorithms to applications*. 728.

Fröhlking, T., Bernetti, M., Calonaci, N., & Bussi, G. (2020). Toward empirical force fields that match experimental observables. *The Journal of Chemical Physics*, *152*(23), 230902. https://doi.org/10.1063/5.0011346

Gapsys, V., Michielssens, S., Seeliger, D., & De Groot, B. L. (2015). pmx: Automated protein structure and topology generation for alchemical perturbations. *Journal of Computational Chemistry*, *36*(5), 348–354. https://doi.org/10.1002/JCC.23804

Garg, A., & Debnath, A. (2025). Thermodynamic origin of fenugreek phytochemical binding to the ASC pyrin domain for inflammation inhibition. *Physical Chemistry Chemical Physics*, *27*(8), 4211–4221. https://doi.org/10.1039/D4CP04644G

Gillmore, J. D., Gane, E., Taubel, J., Kao, J., Fontana, M., Maitland, M. L., Seitzer, J., O'Connell, D., Walsh, K. R., Wood, K., Phillips, J., Xu, Y., Amaral, A., Boyd, A. P., Cehelsky, J. E., McKee, M. D., Schiermeier, A., Harari, O., Murphy, A., … Lebwohl, D. (2021). CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. *New England Journal of Medicine*, *385*(6), 493–502. https://doi.org/10.1056/NEJMOA2107454

Gong, S., Yu, H. H., Johnson, K. A., & Taylor, D. W. (2018). DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell Reports*, *22*(2), 359–371. https://doi.org/10.1016/J.CELREP.2017.12.041

Gouw, A. (2019). The CRISPR Advent of Lulu and Nana. *Https://Doi.Org/10.1080/14746700.2018.1557378*, *17*(1), 9–12. https://doi.org/10.1080/14746700.2018.1557378

Greely, H. T. (2019). CRISPR'd babies: human germline genome editing in the 'He Jiankui affair.' *Journal of Law and the Biosciences*, *6*(1), 111–183. https://doi.org/10.1093/JLB/LSZ010

Guiderdoni, M. N. E., Asante, . M D, Quain, M. D., Ribeiro, P., Ochar, K., Egbadzor, K., Kotey, . D, & A1. (2023). Overview of CRISPR-Cas9 technologies and its application in crop improvement. *International Journal of Genetics and Genomic Science*, *1*(1). https://doi.org/10.58489/IJGGS.006

Guo, M., Ren, K., Zhu, Y., Tang, Z., Wang, Y., Zhang, B., & Huang, Z. (2019a). Structural insights into a high fidelity variant of SpCas9. *Cell Research 2019 29:3*, *29*(3), 183–192. https://doi.org/10.1038/s41422-018-0131-6

Hansen, N., & Van Gunsteren, W. F. (2014). Practical Aspects of Free-Energy Calculations: A Review. *Journal of Chemical Theory and Computation*, *10*(7), 2632–2647. https://doi.org/10.1021/CT500161F

Hassan, M. M., Zhang, Y., Yuan, G., De, K., Chen, J. G., Muchero, W., Tuskan, G. A., Qi, Y., & Yang, X. (2021). Construct design for CRISPR/Cas-based genome editing in plants. *Trends in Plant Science*, *26*(11), 1133–1152. https://doi.org/10.1016/J.TPLANTS.2021.06.015

Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K., & Doudna, J. A. (2010). Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, *329*(5997), 1355–1358. https://doi.org/10.1126/SCIENCE.1192272

Helene, C. (1977). Specific recognition of guanine bases in protein-nucleic acid complexes. *FEBS Letters*, *74*(1), 10–13. https://doi.org/10.1016/0014-5793(77)80740-0

Hess, B. (2007). P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation*, *4*(1), 116–122. https://doi.org/10.1021/CT700200B

Hess, B., Bekker, H., Berendsen, H. J. C., & Fraaije, J. G. E. M. (1997). LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput Chem*, *18*, 14631472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12

Hestenes, M. R., & Stiefel, E. (1952). Methods of Conjugate Gradients for Solving Linear Systems 1. *Journal of Research of the National Bureau of Standards*, *49*(6).

Hibshman, G. N., Bravo, J. P. K., Hooper, M. M., Dangerfield, T. L., Zhang, H., Finkelstein, I. J., Johnson, K. A., & Taylor, D. W. (2024). Unraveling the mechanisms of PAMless DNA interrogation by SpRY-Cas9. *Nature Communications 2024 15:1*, *15*(1), 1–15. https://doi.org/10.1038/s41467-024-47830-3

Hillary, V. E., & Ceasar, S. A. (2023). A Review on the Mechanism and Applications of CRISPR/Cas9/Cas12/Cas13/Cas14 Proteins Utilized for Genome Engineering. *Molecular Biotechnology*, *65*(3), 311–325. https://doi.org/10.1007/S12033-022-00567-0

Hossain, K. A., Kogut, M., Słabonska, J., Sappati, S., Wieczór, M., & Czub, J. (2023). How acidic amino acid residues facilitate DNA target site selection. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(3), e2212501120. https://doi.org/10.1073/PNAS.2212501120

Hossain, K. A., Nierzwicki, L., Orozco, M., Czub, J., & Palermo, G. (2025). Flexibility in PAM Recognition Expands DNA Targeting in xCas9. *ELife*, *13*. https://doi.org/10.7554/ELIFE.102538.2

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell*, *157*(6), 1262. https://doi.org/10.1016/J.CELL.2014.05.010

Hu, J. H., Miller, S. M., Geurts, M. H., Tang, W., Chen, L., Sun, N., Zeina, C. M., Gao, X., Rees, H. A., Lin, Z., & Liu, D. R. (2018). Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature 2018 556:7699*, *556*(7699), 57–63. https://doi.org/10.1038/nature26155

Huang, J., & Mackerell, A. D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, *34*(25), 2135–2145. https://doi.org/10.1002/JCC.23354

Hummer, G., Pratt, L. R., & Garcia, A. E. (1996). Free energy of ionic hydration. *Journal of Physical Chemistry*, *100*(4), 1206–1215. https://doi.org/10.1021/JP951011V

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

Jansen, R., Van Embden, J. D. A., Gaastra, W., & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, *43*(6), 1565–1575. https://doi.org/10.1046/J.1365-2958.2002.02839.X

Jiang, C., Lv, G., Tu, Y., Cheng, X., Duan, Y., Zeng, B., & He, B. (2021). Applications of CRISPR/Cas9 in the Synthesis of Secondary Metabolites in Filamentous Fungi. *Frontiers in Microbiology*, *12*, 164. https://doi.org/10.3389/FMICB.2021.638096

Jiang, F., & Doudna, J. A. (2017). CRISPR–Cas9 Structures and Mechanisms. *Https://Doi.Org/10.1146/Annurev-Biophys-062215-010822*, *46*, 505–529. https://doi.org/10.1146/ANNUREV-BIOPHYS-062215-010822

Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E., & Doudna, J. A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, *351*(6275), 867–871. https://doi.org/10.1126/SCIENCE.AAD8282

Jiang, F., Zhou, K., Ma, L., Gressel, S., & Doudna, J. A. (2015). A Cas9-guide RNA complex preorganized for target DNA recognition. *Science*, *348*(6242), 1477–1481. https://doi.org/10.1126/SCIENCE.AAB1452

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/SCIENCE.1225829

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A. T., Charpentier, E., Nogales, E., & Doudna, J. A. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, *343*(6176). https://doi.org/10.1126/SCIENCE.1247997

Jing, Z., Liu, C., Cheng, S. Y., Qi, R., Walker, B. D., Piquemal, J. P., & Ren, P. (2019). Polarizable force fields for biomolecular simulations: Recent advances and applications. *Annual Review of Biophysics*, *48*, 371. https://doi.org/10.1146/ANNUREV-BIOPHYS-070317-033349

Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, *29*(11), 1859–1865. https://doi.org/10.1002/JCC.20945

Jones, D. L., Leroy, P., Unoson, C., Fange, D., Ćurić, V., Lawson, M. J., & Elf, J. (2017). Kinetics of dCas9 target search in Escherichia coli. *Science*, *357*(6358), 1420–1424. https://doi.org/10.1126/SCIENCE.AAH7084

Jorgensen, W. L., Buckner, J. K., Boudon, S., & Tirado-Rives, J. (1988). Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *The Journal of Chemical Physics*, *89*(6), 3742–3746. https://doi.org/10.1063/1.454895

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1998). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, *79*(2), 926. https://doi.org/10.1063/1.445869

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, *118*(45), 11225–11236. https://doi.org/10.1021/JA9621760

Kang, M., Zuo, Z., Yin, Z., & Gu, J. (2022). Molecular Mechanism of D1135E-Induced Discriminated CRISPR-Cas9 PAM Recognition. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ACS.JCIM.1C01562

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology 2002 9:9*, *9*(9), 646–652. https://doi.org/10.1038/nsb0902-646

Khalak, Y., Tresadern, G., Aldeghi, M., Baumann, H. M., Mobley, D. L., de Groot, B. L., & Gapsys, V. (2021). Alchemical absolute protein–ligand binding free energies for drug design. *Chemical Science*, *12*(41), 13958–13971. https://doi.org/10.1039/D1SC03472C

Kieper, S. N., Almendros, C., & Brouns, S. J. J. (2019). Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. *FEMS Microbiology Letters*, *366*(11). https://doi.org/10.1093/FEMSLE/FNZ129

Kiernan, K. A., & Taylor, D. W. (2025). Visualization of a multi-turnover Cas9 after product release. *Nature Communications 2025 16:1*, *16*(1), 1–11. https://doi.org/10.1038/s41467-025-60668-7

Kim, B., Kim, H. J., & Lee, S. J. (2020). Regulation of Microbial Metabolic Rates Using CRISPR Interference With Expanded PAM Sequences. *Frontiers in Microbiology*, *11*, 503271. https://doi.org/10.3389/FMICB.2020.00282/BIBTEX

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P. W., Li, Z., Peterson, R. T., Yeh, J. R. J., Aryee, M. J., & Joung, J. K. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature 2015 523:7561*, *523*(7561), 481–485. https://doi.org/10.1038/nature14592

Koonin, E. V., Makarova, K. S., & Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Current Opinion in Microbiology*, *37*, 67–78. https://doi.org/10.1016/J.MIB.2017.05.008

Krimsky, S. (2019). Ten ways in which He Jiankui violated ethics. *Nature Biotechnology 2019 37:1*, *37*(1), 19–20. https://doi.org/10.1038/nbt.4337

Kumar, A., Basu, D., & Satpati, P. (2017). Structure-Based Energetics of Stop Codon Recognition by Eukaryotic Release Factor. *Journal of Chemical Information and Modeling*, *57*(9), 2321–2328. https://doi.org/10.1021/ACS.JCIM.7B00340

Lander, E. S. (2016). The Heroes of CRISPR. *Cell*, *164*(1–2), 18–28. https://doi.org/10.1016/J.CELL.2015.12.041

Lapelosa, M., Gallicchio, E., & Levy, R. M. (2012). Conformational Transitions and Convergence of Absolute Binding Free Energy Calculations. *Journal of Chemical Theory and Computation*, *8*(1), 47. https://doi.org/10.1021/CT200684B

Le Rhun, A., Escalera-Maurer, A., Bratovič, M., & Charpentier, E. (2019). CRISPR-Cas in Streptococcus pyogenes. *Https://Doi.Org/10.1080/15476286.2019.1582974*, *16*(4), 380–389. https://doi.org/10.1080/15476286.2019.1582974

Levrel, L., & Maggs, A. C. (2008). Boundary conditions in local electrostatics algorithms. *Journal of Chemical Physics*, *128*(21), 214103. https://doi.org/10.1063/1.2918365/959057

Li, P., & Merz, K. M. (2013). Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions. *Journal of Chemical Theory and Computation*, *10*(1), 289–297. https://doi.org/10.1021/CT400751U

Liao, B., Chen, X., Zhou, X., Zhou, Y., Shi, Y., Ye, X., Liao, M., Zhou, Z., Cheng, L., & Ren, B. (2021). Applications of CRISPR/Cas gene-editing technology in yeast and fungi. *Archives of Microbiology 2021 204:1*, *204*(1), 1–14. https://doi.org/10.1007/S00203-021-02723-7

Lin, Y. L., Aleksandrov, A., Simonson, T., & Roux, B. (2014). An overview of electrostatic free energy computations for solutions and proteins. *Journal of Chemical Theory and Computation*, *10*(7), 2690–2709. https://doi.org/10.1021/CT500195P

Lind, C., Esguerra, M., Jespers, W., Satpati, P., Gutierrez-de-Terán, H., & Åqvist, J. (2019). Free energy calculations of RNA interactions. *Methods*, *162–163*, 85–95. https://doi.org/10.1016/J.YMETH.2019.02.014

Liu, P., Dehez, F., Cai, W., & Chipot, C. (2012). A toolkit for the analysis of free-energy perturbation calculations. *Journal of Chemical Theory and Computation*, *8*(8), 2606–2616. https://doi.org/10.1021/CT300242F

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., … Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, *102*(18), 3586–3616. https://doi.org/10.1021/JP973084F

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., Van Der Oost, J., & Koonin, E. V. (2011). Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology 2011 9:6*, *9*(6), 467–477. https://doi.org/10.1038/nrmicro2577

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R.

M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., Van Der Oost, J., … Koonin, E. V. (2015). An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology 2015 13:11*, *13*(11), 722–736. https://doi.org/10.1038/nrmicro3569

Mani, I. (2021). CRISPR-Cas9 for treating hereditary diseases. *Progress in Molecular Biology and Translational Science*, *181*, 165–183. https://doi.org/10.1016/BS.PMBTS.2021.01.017

Marraffini, L. A. (2015). CRISPR-Cas immunity in prokaryotes. *Nature 2015 526:7571*, *526*(7571), 55–61. https://doi.org/10.1038/nature15386

Martyna, G. J., Tobias, D. J., & Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, *101*(5), 4177–4189. https://doi.org/10.1063/1.467468

McGinn, J., & Marraffini, L. A. (2018). Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nature Reviews Microbiology 2018 17:1*, *17*(1), 7–12. https://doi.org/10.1038/s41579-018-0071-7

Mey, A. S. J. S., Allen, B. K., Bruce Macdonald, H. E., Chodera, J. D., Hahn, D. F., Kuhn, M., Michel, J., Mobley, D. L., Naden, L. N., Prasad, S., Rizzi, A., Scheen, J., Shirts, M. R., Tresadern, G., & Xu, H. (2020). Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living Journal of Computational Molecular Science*, *2*(1), 18378. https://doi.org/10.33011/LIVECOMS.2.1.18378

Meza, J. C. (2010). Steepest descent. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(6), 719–722. https://doi.org/10.1002/WICS.117

Mitchell, B. P., Hsu, R. V., Medrano, M. A., Zewde, N. T., Narkhede, Y. B., & Palermo, G. (2020). Spontaneous Embedding of DNA Mismatches Within the RNA:DNA Hybrid of CRISPR-Cas9. *Frontiers in Molecular Biosciences*, *7*, 39. https://doi.org/10.3389/FMOLB.2020.00039

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*, *60*(2), 174–182. https://doi.org/10.1007/S00239-004-0046-3

Molani, F., & Cho, A. E. (2024). Accurate protein-ligand binding free energy estimation using QM/MM on multi-conformers predicted from classical mining minima. *Communications Chemistry 2024 7:1*, *7*(1), 1–10. https://doi.org/10.1038/s42004-024-01328-7

Muegge, I., & Hu, Y. (2023). Recent Advances in Alchemical Binding Free Energy Calculations for Drug Discovery. *ACS Medicinal Chemistry Letters*, *14*(3), 244–250. https://doi.org/10.1021/ACSMEDCHEMLETT.2C00541

Nakata, A., Amemura, M., & Makino, K. (1989). Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. *Journal of Bacteriology*, *171*(6), 3553–3556. https://doi.org/10.1128/JB.171.6.3553-3556.1989

Nguyen, T. M., Lu, C. A., & Huang, L. F. (2022). Applications of CRISPR/Cas9 in a rice protein expression system via an intron-targeted insertion approach. *Plant Science*, *315*, 111132. https://doi.org/10.1016/J.PLANTSCI.2021.111132

Nierzwicki, Ł., Arantes, P. R., Saha, A., & Palermo, G. (2021). Establishing the allosteric mechanism in CRISPR-Cas9. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, *11*(3). https://doi.org/10.1002/WCMS.1503

Nierzwicki, Ł., East, K. W., Binz, J. M., Hsu, R. V., Ahsan, M., Arantes, P. R., Skeens, E., Pacesa, M., Jinek, M., Lisi, G. P., & Palermo, G. (2022). Principles of target DNA cleavage and the role of Mg2+ in the catalysis of CRISPR–Cas9. *Nature Catalysis*, *5*(10), 912. https://doi.org/10.1038/S41929-022-00848-6

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F., & Nureki, O. (2014). Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell*, *156*(5), 935–949. https://doi.org/10.1016/J.CELL.2014.02.001

Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O. O., Gootenberg, J. S., Mori, H., Oura, S., Holmes, B., Tanaka, M., Seki, M., Hirano, H., Aburatani, H., Ishitani, R., Ikawa, M., Yachie, N., … Nureki, O. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science*, *361*(6408), 1259–1262. https://doi.org/10.1126/SCIENCE.AAS9129

Nosé, S., & Klein, M. L. (1983). Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, *50*(5), 1055–1076. https://doi.org/10.1080/00268978300102851

Pacesa, M., Loeff, L., Querques, I., Muckenfuss, L. M., Sawicka, M., & Jinek, M. (2022). R-loop formation and conformational activation mechanisms of Cas9. *Nature 2022 609:7925*, *609*(7925), 191–196. https://doi.org/10.1038/s41586-022-05114-0

Palermo, G. (2019). Structure and Dynamics of the CRISPR-Cas9 Catalytic Complex. *Journal of Chemical Information and Modeling*, *59*(5), 2394–2406. https://doi.org/10.1021/ACS.JCIM.8B00988

Palermo, G., Miao, Y., Walker, R. C., Jinek, M., & McCammon, J. A. (2016). Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations. *ACS Central Science*, *2*(10), 756–763. https://doi.org/10.1021/ACSCENTSCI.6B00218

Palermo, G., Miao, Y., Walker, R. C., Jinek, M., & McCammon, J. A. (2017). CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(28), 7260–7265. https://doi.org/10.1073/PNAS.1707645114

Palermo, G., Ricci, C. G., Fernando, A., Basak, R., Jinek, M., Rivalta, I., Batista, V. S., & McCammon, J. A. (2017). Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *Journal of the American Chemical Society*, *139*(45), 16028–16031. https://doi.org/10.1021/JACS.7B05313

Panteva, M. T., Giambaşu, G. M., & York, D. M. (2015). Comparison of structural, thermodynamic, kinetic and mass transport properties of Mg(2+) ion models commonly used in biomolecular simulations. *Journal of Computational Chemistry*, *36*(13), 970–982. https://doi.org/10.1002/JCC.23881

Park, S. J., Kern, N., Brown, T., Lee, J., & Im, W. (2023). CHARMM-GUI PDB Manipulator: Various PDB Structural Modifications for Biomolecular Modeling and Simulation. *Journal of Molecular Biology*, *435*(14), 167995. https://doi.org/10.1016/J.JMB.2023.167995

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, *52*(12), 7182–7190. https://doi.org/10.1063/1.328693

Pattan, V., Kashyap, R., Bansal, V., Candula, N., Koritala, T., & Surani, S. (2021). Genomics in medicine: A new era in medicine. *World Journal of Methodology*, *11*(5), 231. https://doi.org/10.5662/WJM.V11.I5.231

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, *26*(16), 1781–1802. https://doi.org/10.1002/JCC.20289

Pickar-Oliver, A., & Gersbach, C. A. (2019). The next generation of CRISPR–Cas technologies and applications. *Nature Reviews. Molecular Cell Biology*, *20*(8), 490. https://doi.org/10.1038/S41580-019-0131-5

Pohorille, A., Jarzynski, C., & Chipot, C. (2010). Good practices in free-energy calculations. *Journal of Physical Chemistry B*, *114*(32), 10235–10253. https://doi.org/10.1021/JP102971X

Ponder, J. W., & Case, D. A. (2003). Force fields for protein simulations. *Advances in Protein Chemistry*, *66*, 27–85. https://doi.org/10.1016/S0065-3233(03)66002-X

Pourcel, C., Salvignol, G., & Vergnaud, G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, *151*(3), 653–663. https://doi.org/10.1099/MIC.0.27437-0

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J. P., Couvin, D., Toffano-Nioche, C., & Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Research*, *48*(D1), D535–D544. https://doi.org/10.1093/NAR/GKZ915

Rahman, A., & Stillinger, F. H. (1971). Molecular Dynamics Study of Liquid Water. *The Journal of Chemical Physics*, *55*(7), 3336–3359. https://doi.org/10.1063/1.1676585

Rahman, S., Ikram, A. R., Azeem, F., Tahir ul Qamar, M., Shaheen, T., & Mehboob-ur-Rahman. (2024). Precision Genome Editing with CRISPR-Cas9. *Methods in Molecular Biology (Clifton, N.J.)*, *2788*, 355–372. https://doi.org/10.1007/978-1-0716-3782-1_21

Raper, A. T., Stephenson, A. A., & Suo, Z. (2018). Functional Insights Revealed by the Kinetic Mechanism of CRISPR/Cas9. *Journal of the American Chemical Society*, *140*(8), 2971–2984. https://doi.org/10.1021/JACS.7B13047

Rath, D., Amlinger, L., Rath, A., & Lundgren, M. (2015). The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie*, *117*, 119–128. https://doi.org/10.1016/J.BIOCHI.2015.03.025

Ray, A., & Di Felice, R. (2020). Protein-Mutation-Induced Conformational Changes of the DNA and Nuclease Domain in CRISPR/Cas9 Systems by Molecular Dynamics Simulations. *Journal of Physical Chemistry B*, *124*(11), 2168–2179. https://doi.org/10.1021/ACS.JPCB.9B07722

Ray, A., Felice, R. Di, Felice, R. Di, & Felice, R. Di. (2019). Molecular Simulations have Boosted Knowledge of CRISPR/Cas9: A Review. *Journal of Self-Assembly and Molecular Electronics (SAME)*, *7*(1), 45–72. https://doi.org/10.13052/JSAME2245-4551.7.003

Ren, B., Liu, L., Li, S., Kuang, Y., Wang, J., Zhang, D., Zhou, X., Lin, H., & Zhou, H. (2019). Cas9-NG Greatly Expands the Targeting Scope of the Genome-Editing Toolkit by Recognizing NG and Other Atypical PAMs in Rice. *Molecular Plant*, *12*(7), 1015–1026. https://doi.org/10.1016/J.MOLP.2019.03.010

Ricci, C. G., Chen, J. S., Miao, Y., Jinek, M., Doudna, J. A., McCammon, J. A., & Palermo, G. (2019). Deciphering Off-Target Effects in CRISPR-Cas9 through Accelerated Molecular Dynamics. *ACS Central Science*, *5*(4), 651–662. https://doi.org/10.1021/ACSCENTSCI.9B00020

Riniker, S. (2018). Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *Journal of Chemical Information and Modeling*, *58*(3), 565–578. https://doi.org/10.1021/ACS.JCIM.8B00042

Roberts, J. E., & Schnitker, J. (1995). Boundary conditions in simulations of aqueous ionic solutions: A systematic study. *Journal of Physical Chemistry*, *99*(4), 1322–1331. https://doi.org/10.1021/J100004A037

Rocklin, G. J., Mobley, D. L., Dill, K. A., & Hünenberger, P. H. (2013). Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *Journal of Chemical Physics*, *139*(18), 184103. https://doi.org/10.1063/1.4826261/317441

Sadeqi Nezhad, M., Yazdanifar, M., Abdollahpour-Alitappeh, M., Sattari, A., Seifalian, A., & Bagheri, N. (2021). Strengthening the CAR-T cell therapeutic application using CRISPR/Cas9 technology. *Biotechnology and Bioengineering*, *118*(10), 3691–3705. https://doi.org/10.1002/BIT.27882

Satpati, P., & Åqvist, J. (2014). Why base tautomerization does not cause errors in mRNA decoding on the ribosome. *Nucleic Acids Research*, *42*(20), 12876–12884. https://doi.org/10.1093/NAR/GKU1044

Satpati, P., Sund, J., & Åqvist, J. (2014). Structure-based energetics of mRNA decoding on the ribosome. *Biochemistry*, *53*(10), 1714–1722. https://doi.org/10.1021/BI5000355

Scott, W. R. P., Hu, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Kru, P., & Van Gunsteren, W. F. (1999). *The GROMOS Biomolecular Simulation Program Package*. https://doi.org/10.1021/jp984217f

Seok, H., Deng, R., Cowan, D. B., & Wang, D. Z. (2021). Application of CRISPR-Cas9 gene editing for congenital heart disease. *Clinical and Experimental Pediatrics*, *64*(6), 269. https://doi.org/10.3345/CEP.2020.02096

Sharma, V., Panwar, A., Gupta, G. K., & Sharma, A. K. (2022). Molecular docking and MD: mimicking the real biological process. *Physical Sciences Reviews*, *0*(0). https://doi.org/10.1515/PSR-2018-0164

Shell, M. S., Panagiotopoulos, A., & Pohorille, A. (2007). Methods based on probability distributions and histograms. *Springer Series in Chemical Physics*, *86*, 77–118. https://doi.org/10.1007/978-3-540-38448-9_3

Shobana, S., Roux, B., & Andersen, O. S. (2000). Free Energy Simulations: Thermodynamic Reversibility and Variability. *The Journal of Physical Chemistry B*, *104*(21), 5179–5190. https://doi.org/10.1021/JP994193S

Shukla, S., Kumar, A., Das, D., & Satpati, P. (2020). Principle of DNA recognition by sporulation-regulatory protein (Spo0A) in Bacillus subtilis. *Journal of Biomolecular Structure & Dynamics*, *38*(17), 5186–5194. https://doi.org/10.1080/07391102.2019.1696890

Simonson, T., & Satpati, P. (2013). Simulating GTP:Mg and GDP:Mg with a simple force field: a structural and thermodynamic analysis. *Journal of Computational Chemistry*, *34*(10), 836–846. https://doi.org/10.1002/JCC.23207

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

Singh, N., & Warshel, A. (2010). Absolute Binding Free Energy Calculations: On the Accuracy of Computational Scoring of Protein-ligand Interactions. *Proteins*, *78*(7), 1705. https://doi.org/10.1002/PROT.22687

Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X., & Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. *Science*, *351*(6268), 84–88. https://doi.org/10.1126/SCIENCE.AAD5227

Song, B., Yang, S., Hwang, G. H., Yu, J., & Bae, S. (2021). Analysis of nhej-based dna repair after crispr-mediated dna cleavage. *International Journal of Molecular Sciences*, *22*(12). https://doi.org/10.3390/IJMS22126397

Sorolla, A., Parisi, E., Sorolla, M. A., Marqués, M., & Porcel, J. M. (2022). Applications of CRISPR technology to lung cancer research. *European Respiratory Journal*, *59*(1), 2102610. https://doi.org/10.1183/13993003.02610-2021

Steinbrecher, T., & Labahn, A. (2010). Towards accurate free energy calculations in ligand protein-binding studies. *Current Medicinal Chemistry*, *17*(8), 767–785. https://doi.org/10.2174/092986710790514453

Sternberg, S. H., Lafrance, B., Kaplan, M., & Doudna, J. A. (2015). Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature 2015 527:7576*, *527*(7576), 110–113. https://doi.org/10.1038/nature15544

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., & Doudna, J. A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature 2014 507:7490*, *507*(7490), 62–67. https://doi.org/10.1038/nature13011

Swope, W. C., Andersen, H. C., Berens, P. H., & Wilson, K. R. (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. https://doi.org/10.1063/1.442716

Tang, L., Zeng, Y., Du, H., Gong, M., Peng, J., Zhang, B., Lei, M., Zhao, F., Wang, W., Li, X., & Liu, J. (2017). CRISPR/Cas9-mediated gene editing in human zygotes using Cas9 protein. *Molecular Genetics and Genomics*, *292*(3), 525–533. https://doi.org/10.1007/S00438-017-1299-Z

Taylor, H. N., Laderman, E., Armbrust, M., Hallmark, T., Keiser, D., Bondy-Denomy, J., & Jackson, R. N. (2021). Positioning Diverse Type IV Structures and Functions Within Class 1 CRISPR-Cas Systems. *Frontiers in Microbiology*, *12*, 1236. https://doi.org/10.3389/FMICB.2021.671522

Tiruneh G/Medhin, M., Abebe, E. C., Sisay, T., Berhane, N., Snr, T. B., & Dejenie, T. A. (2021). Current Applications and Future Perspectives of CRISPR-Cas9 for the Treatment of Lung Cancer. *Biologics : Targets & Therapy*, *15*, 199. https://doi.org/10.2147/BTT.S310312

Torella, R., Moroni, E., Caselle, M., Morra, G., & Colombo, G. (2010). Investigating dynamic and energetic determinants of protein nucleic acid recognition: Analysis of the zinc finger zif268-DNA complexes. *BMC Structural Biology*, *10*(1), 1–18. https://doi.org/10.1186/1472-6807-10-42

Trobro, S., & Åqvist, J. (2005). Mechanism of peptide bond synthesis on the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(35), 12395–12400. https://doi.org/10.1073/PNAS.0504043102

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, *26*(16), 1701–1718. https://doi.org/10.1002/JCC.20291

Van Heesch, T., Bolhuis, P. G., & Vreede, J. (2023). Decoding dissociation of sequence-specific protein-DNA complexes with non-equilibrium simulations. *Nucleic Acids Research*, *51*(22), 12150–12160. https://doi.org/10.1093/NAR/GKAD1014

Vanommeslaeghe, K., & Mackerell, A. D. (2014). CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta*, *1850*(5), 861. https://doi.org/10.1016/J.BBAGEN.2014.08.004

Wahab, A., Abaseen, N., Hayat, M., Khan, B., & Luqman, M. (2022). Advances in understanding the DNA-repair mechanism activated by CRISPR/Cas9. *International Journal of Biology Sciences*, *4*(2), 01–10. https://doi.org/10.33545/26649926.2022.V4.I2A.68

Walton, R. T., Christie, K. A., Whittaker, M. N., & Kleinstiver, B. P. (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science*, *368*(6488), 290–296. https://doi.org/10.1126/SCIENCE.ABA8853

Wang, J., Arantes, P. R., Ahsan, M., Sinha, S., Kyro, G. W., Maschietto, F., Allen, B., Skeens, E., Lisi, G. P., Batista, V. S., & Palermo, G. (2023). Twisting and swiveling domain motions in Cas9 to recognize target DNA duplexes, make double-strand breaks, and release cleaved duplexes. *Frontiers in Molecular Biosciences*, *9*, 1072733. https://doi.org/10.3389/FMOLB.2022.1072733

Wang, W., Liang, Z., Ma, P., Zhao, Q., Dai, M., Zhu, J., Han, X., Xu, H., Chang, Q., & Zhen, Y. (2021). Application of CRISPR/Cas9 System to Reverse ABC-Mediated Multidrug Resistance. *Bioconjugate Chemistry*, *32*(1), 73–81. https://doi.org/10.1021/ACS.BIOCONJCHEM.0C00627

Warren, R. M., Streicher, E. M., Sampson, S. L., Van der Spuy, G. D., Richardson, M., Nguyen, D., Behr, M. A., Victor, T. C., & Van Helden, P. D. (2002). Microevolution of the direct repeat region of Mycobacterium tuberculosis: Implications for interpretation of spoligotyping data. *Journal of Clinical Microbiology*, *40*(12), 4457–4465. https://doi.org/10.1128/JCM.40.12.4457-4465.2002

Wellhausen, N., Agarwal, S., Rommel, P. C., Gill, S. I., & June, C. H. (2021). Better living through chemistry: CRISPR/Cas engineered T cells for cancer immunotherapy. *Current Opinion in Immunology*, *74*, 76. https://doi.org/10.1016/J.COI.2021.10.008

Xue, C., Yang, C., Zhou, Z., Sun, X., Ju, H., Yue, X., & Rao, S. (2023). *PAMless SpRY recognizes a non-PAM region for efficient targeting*. https://doi.org/10.21203/RS.3.RS-3177819/V1

Yamaguchi, H., Siebers, J. G., Furukawa, A., Otagiri, N., Osman, R., Cherubini, R., Goodhead, D. T., Menzel, H. G., & Ottolenghi, A. (2002). Molecular dynamics simulation of a DNA containing a single strand break. *Radiation Protection Dosimetry*, *99*(1–4), 103–108. https://doi.org/10.1093/OXFORDJOURNALS.RPD.A006737

Yin, H., Song, C. Q., Dorkin, J. R., Zhu, L. J., Li, Y., Wu, Q., Park, A., Yang, J., Suresh, S., Bizhanova, A., Gupta, A., Bolukbasi, M. F., Walsh, S., Bogorad, R. L., Gao, G., Weng, Z., Dong, Y., Koteliansky, V., Wolfe, S. A., … Anderson, D. G. (2016). Therapeutic genome editing by combined viral and non-viral delivery of CRISPR system components in vivo. *Nature Biotechnology 2016 34:3*, *34*(3), 328–333. https://doi.org/10.1038/nbt.3471

Yoo, J., & Aksimentiev, A. (2012). Competitive Binding of Cations to Duplex DNA Revealed through Molecular Dynamics Simulations. *Journal of Physical Chemistry B*, *116*(43), 12946–12954. https://doi.org/10.1021/JP306598Y

Zeng, X., Chugh, J., Casiano-Negroni, A., Al-Hashimi, H. M., & Brooks, C. L. (2014). Flipping of the ribosomal A-site adenines provides a basis for tRNA selection. *Journal of Molecular Biology*, *426*(19), 3201. https://doi.org/10.1016/J.JMB.2014.04.029

Zeng, Y., Cui, Y., Zhang, Y., Zhang, Y., Liang, M., Chen, H., Lan, J., Song, G., & Lou, J. (2018). The initiation, propagation and dynamics of CRISPR-SpyCas9 R-loop complex. *Nucleic Acids Research*, *46*(1), 350–361. https://doi.org/10.1093/NAR/GKX1117

Zheng, W. (2017). Probing the structural dynamics of the CRISPR-Cas9 RNA-guided DNA-cleavage system by coarse-grained modeling. *Proteins: Structure, Function, and Bioinformatics*, *85*(2), 342–353. https://doi.org/10.1002/PROT.25229

Zhu, H., Wang, H., Wang, L., & Zheng, Z. (2024). CRISPR/Cas9-based genome engineering in the filamentous fungus Rhizopus oryzae and its application to L-lactic acid production. *Biotechnology Journal*, *19*(9). https://doi.org/10.1002/BIOT.202400309

Zhu, X., Clarke, R., Puppala, A. K., Chittori, S., Merk, A., Merrill, B. J., Simonović, M., & Subramaniam, S. (2019). Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nature Structural & Molecular Biology 2019 26:8*, *26*(8), 679–685. https://doi.org/10.1038/s41594-019-0258-2

Zuo, Z., & Liu, J. (2016). Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations. *Scientific Reports 2016 6:1*, *6*(1), 1–9. https://doi.org/10.1038/srep37584

Zuo, Z., & Liu, J. (2017). Structure and Dynamics of Cas9 HNH Domain Catalytic State. *Scientific Reports 2017 7:1*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-17578-6

Zuo, Z., & Liu, J. (2020). Allosteric regulation of CRISPR-Cas9 for DNA-targeting and cleavage. *Current Opinion in Structural Biology*, *62*, 166–174. https://doi.org/10.1016/J.SBI.2020.01.013

Zwanzig, R. W. (2004). High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, *22*(8), 1420. https://doi.org/10.1063/1.1740409

Appendices

Chapter 2

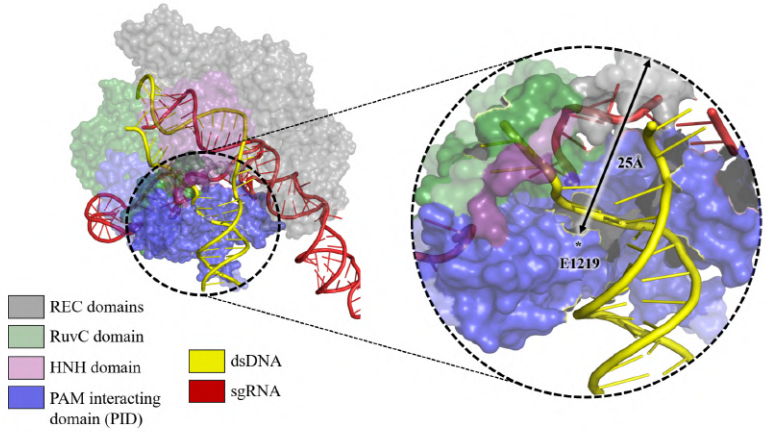


Figure A2.1. Complete X-ray structure of precatalytic *Sp*Cas9 (color-coded surface represents different domains, left). The zoomed in view of 25 Å spherically truncated region centered at E1219 residue is marked with broken circle (right), color-coding highlighted the extent of domains included for simulation.

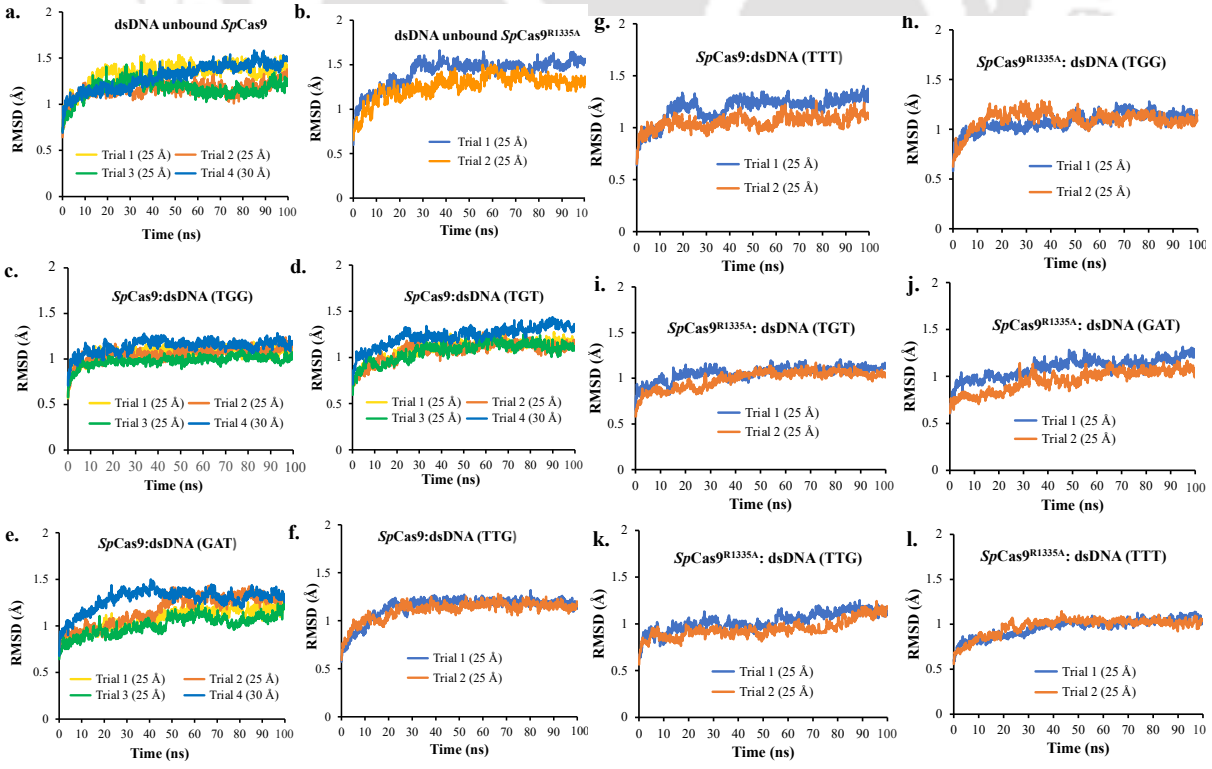


Figure A2.2. Root mean squared deviation (RMSD) versus time plots for 100 ns simulations of multiple independent trials of *Sp*Cas9 and *Sp*Cas9^{R1335A} bound to different DNA substrates and DNA unbound form. The plateau in the RMSD versus time plot depicts the structural convergence. The RMSD values were calculated for the unrestrained protein-heavy atoms (inner region of the spherically truncated systems).

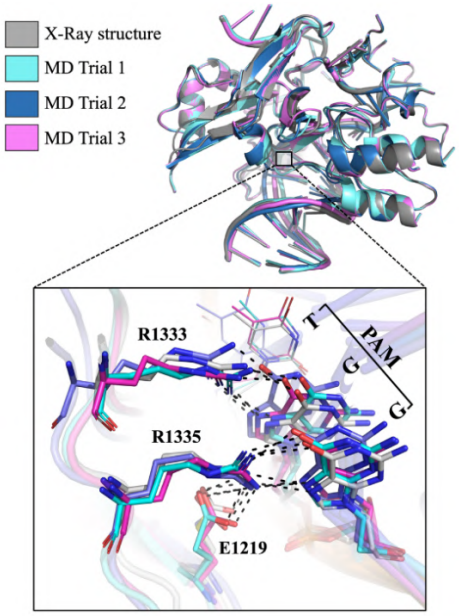


Figure A2.3. Overlay of X-ray (grey, PDB 5F9R; precatalytic *Sp*Cas9) and final structures from three independent MD replicas (colored: cyan, blue, and pink for trials 1, 2, and 3, respectively). Zoomed in view of the *Sp*Cas9:PAM interaction (black outlined box). Key amino acid residues (R1333, R1335, and E1219) and the nucleotides of the cognate PAM sequence (TGG) are shown in the sticks.

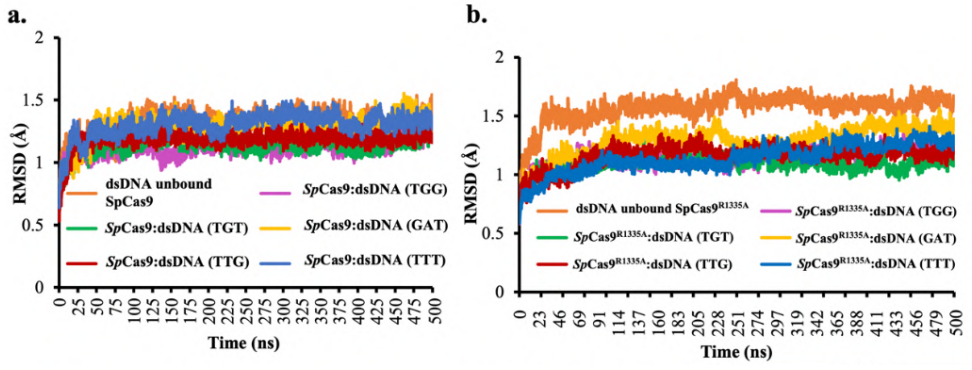


Figure A2.4. (a, b) Heavy atom "RMSD" of the biomolecule as a function of time after 500 ns of extended simulations.

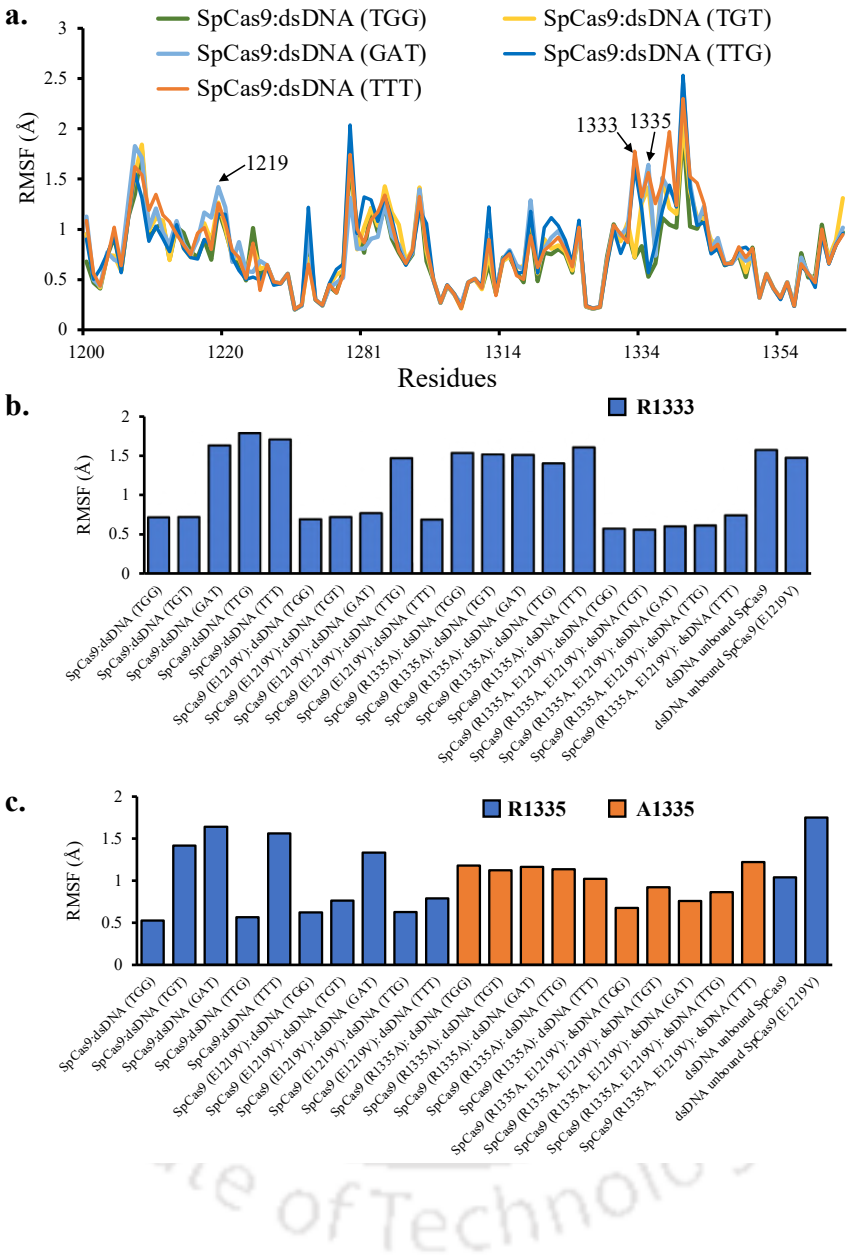


Figure A2.5. **(a)** Root Mean Square Fluctuation (RMSF) of the PAM interacting (PI) domain residues (residue 1200-1353) of *Sp*Cas9 in complex with the various DNA substrates differing in their PAM sequences. **(b)** RMSF of the R1333 residue of *Sp*Cas9 in complex with various PAM containing dsDNA and in dsDNA unbound state, **(c)** RMSF of the R1335/A1335 residues of *Sp*Cas9 in complex with various PAM containing dsDNA and in dsDNA unbound state. The last 50 ns of the production dynamics were used to estimate the RMSF.

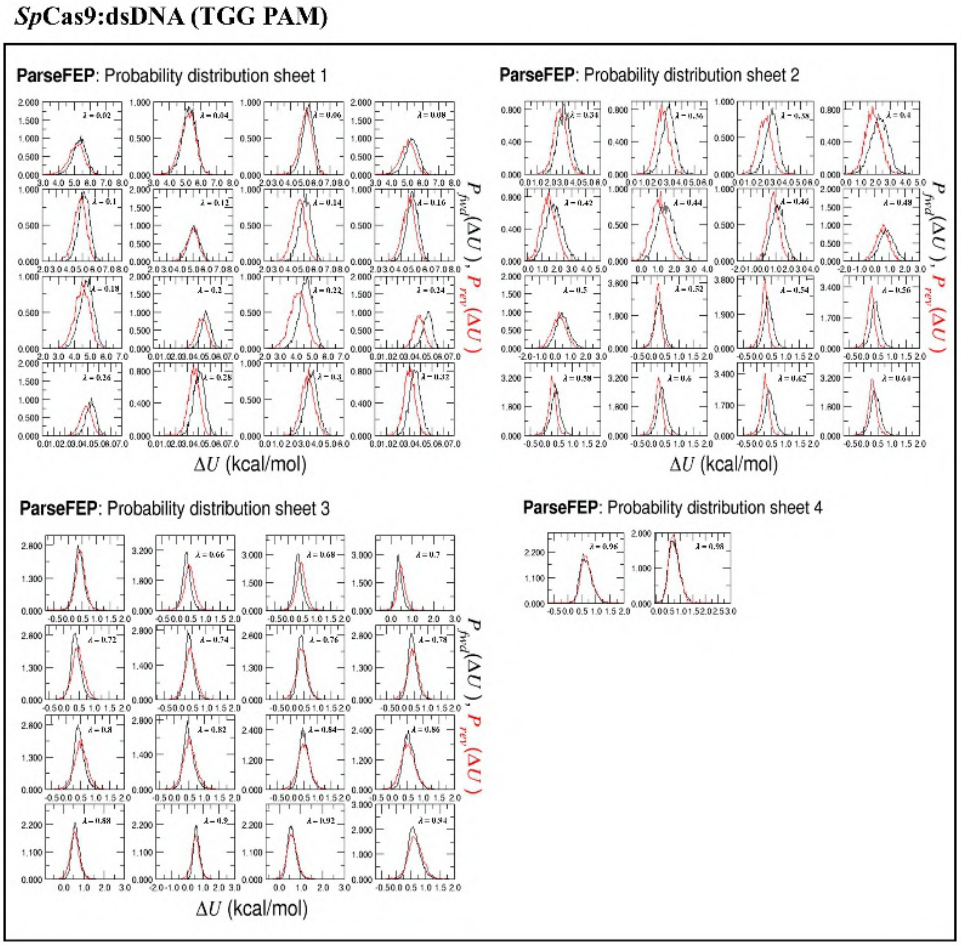


Figure A2.6. Probability distribution function (Y-axis) demonstrating the forward (black) and reverse (red) transformations for all 49 intermediate states in the 51 λ window free energy simulations for *Sp*Cas9:dsDNA (TGG PAM). The X-axis denotes potential energy differences (in kcal/mole) between two neighboring windows. The probability distributions for the forward (black) and reverse (red) significantly overlapped, ensuring reversibility.

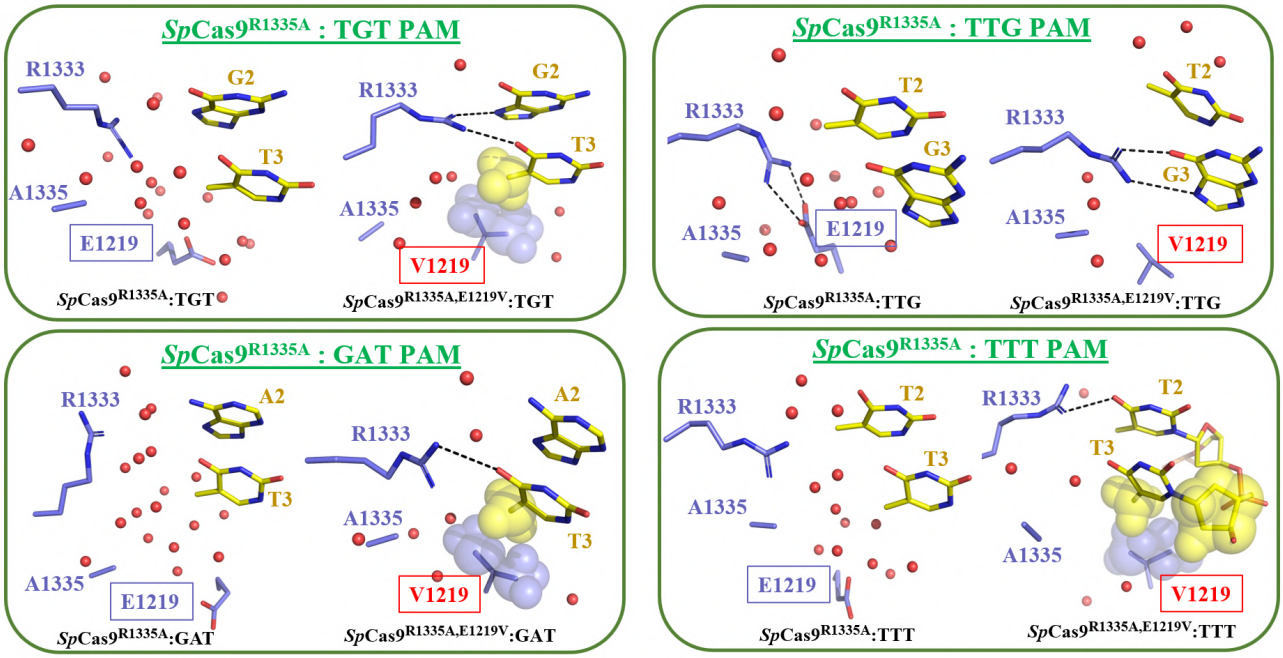


Figure A2.7. Structural comparison of PAM binding pocket in the *Sp*Cas9^{R1335A} (left) and *Sp*Cas9^{R1335A,E1219V} (right) pre-catalytic complex bound to different non-canonical PAMs (TGT/GAT/TTG/TTT).

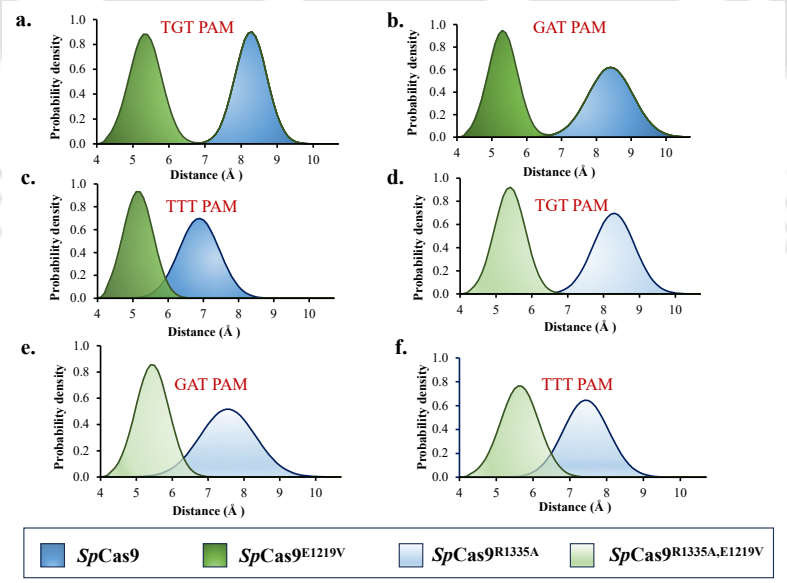


Figure A2.8. Probability distribution plots: The distance between the Cβ atom of E1219/V1219 and the C5 atom of T3 (3rd Thymine of PAM) in *Sp*Cas9 bound to different non-canonical PAM. In all the situations, the Cβ atom of V1219 is closer to the methyl group of 3rd thymine than Cβ atom of E1219.

Table A2.1 Details of various models of the initial setup of the simulation box.

Simulation Model	**Radius of Spherically Truncated model (Å)**	**Box Dimensions (Å^3)**	**Total number of atoms**	**Number of waters/Na**	**Number of trials**
*Sp*Cas9:dsDNA (TGG)	25	78.6 × 82.1 × 77.7	45866	13478/28	3
*Sp*Cas9:dsDNA (TGG)	30	86.4 × 91.9 × 86.4	63112	18461/41	1
*Sp*Cas9:dsDNA (TGG)	Full system	133.2 × 123.7 × 122.8	210509	60886/133	1
*Sp*Cas9:dsDNA (TGT)	25	78.6 × 82.1 × 77.7	45858	13475/28	3
*Sp*Cas9:dsDNA (TGT)	30	86.6 × 91.9 × 86	62894	18388/41	1
*Sp*Cas9:dsDNA (TGT)	Full system	134 × 123.2 × 136.9	211200	61116/133	1
*Sp*Cas9:dsDNA (GAT)	25	78.6 × 82.1 × 77.7	45861	13476/28	3
*Sp*Cas9:dsDNA (GAT)	30	86.6 × 91.8 × 86	62756	18372/41	1
*Sp*Cas9:dsDNA (TTG)	25	79.0 × 82.1 × 76.1	45102	13223/28	2
*Sp*Cas9:dsDNA (TTT)	25	79.0 × 82.1 × 76.1	45103	13222/28	2
dsDNA unbound *Sp*Cas9	25	77.8 × 81.5 × 73.6	42584	12779/6	3
dsDNA unbound *Sp*Cas9	30	78.9 × 88.7 × 86.5	55579	16483/8	1
dsDNA unbound *Sp*Cas9	Full system	141.7 × 115.8 × 123.3	188402	54516/43	1
*Sp*Cas9^{R1335A}: dsDNA (TGG)	25	78.6 × 82.1 × 77.7	45856	13481/29	2
*Sp*Cas9^{R1335A}: dsDNA (TGT)	25	78.6 × 82.1 × 77.7	45854	13478/29	2
*Sp*Cas9^{R1335A}: dsDNA (GAT)	25	78.6 × 82.1 × 77.7	45848	13476/29	2
*Sp*Cas9^{R1335A}: dsDNA (TTG)	25	79.0 × 82.1 × 76.1	45098	13226/29	2
*Sp*Cas9^{R1335A}: dsDNA (TTT)	25	79.0 × 82.1 × 76.1	45084	13221/29	2
dsDNA unbound *Sp*Cas9^{R1335A}	25	75.4 × 82.1 × 73.4	41675	12635/6	2

Table A2.2. Alchemical free energy calculations associated with E1219 →V1219 transformation. Standard errors were reported after '±' symbol.

System (Alchemical transformation: E1219 → V1219**)**	**Trial**	**Truncation sphere size** (Å)	**Number of windows × ns per window**	**Run length** (ns) (Fwd + Back)	**ΔG (FEP)** (kcal/mole)		**ΔG (BAR estimator)** (kcal/mole)	**ΔG_{comp} or ΔG_{free}** (kcal/mole)
					Forward (E→V, **ΔG^F**)	**Backward** (V→E, **ΔG^R**)		
***Sp*Cas9:dsDNA (TGG)**	1	25	51 × 10	1000	96.45 ± 0.12	-95.52 ± 0.08	96.04 ± 0.08	96.07 ± 0.13
	2		51 × 5	500	96.26 ± 0.19	-95.44 ± 0.12	95.71 ± 0.12	
	3		51 × 5	500	96.24 ± 0.18	-96.84 ± 0.11	96.20 ± 0.11	
	4	30	51 × 5	500	96.64 ± 0.14	96.25 ± 0.16	96.33 ± 0.10	
	5	Full system	51 × 3	300	97.41 ± 0.19	96.20 ± 0.18	96.71 ± 0.13	96.71 ± 0.13

***Sp*Cas9:dsDNA (TGT)**	1	25	51 × 10	1000	92.64 ± 0.12	-92.03 ± 0.09	92.18 ± 0.08	92.10 ± 0.10
	2		51 × 5	500	92.41 ± 0.14	-92.06 ± 0.10	92.13 ± 0.08	
	3		51 × 5	500	92.52 ± 0.10	-92.26 ± 0.11	92.23 ± 0.07	
	4	30	51 × 5	500	92.66 ± 0.12	-91.36 ± 0.20	91.85 ± 0.13	
	5	Full system	51 × 3	300	93.54 ± 0.16	-92.08 ± 0.16	92.72 ± 0.12	92.92 ± 0.12
***Sp*Cas9:dsDNA (GAT)**	1	25	51 × 10	1000	94.54 ± 0.12	-94.08 ± 0.08	94.16 ± 0.06	93.75 ± 0.17
	2		51 × 5	500	94.05 ± 0.10	-93.35 ± 0.14	93.44 ± 0.08	
	3		51 × 5	500	94.07 ± 0.12	-94.01 ± 0.09	93.91 ± 0.09	
	4	30	51 × 5	500	93.86 ± 0.10	-93.63 ± 0.15	93.52 ± 0.08	
***Sp*Cas9:dsDNA (TTG)**	1	25	51 × 5	500	96.46 ± 0.16	-95.33 ± 0.15	95.93 ± 0.13	96.06 ± 0.13
	2		51 × 5	500	97.02 ± 0.14	-95.26 ± 0.09	96.19 ± 0.10	
***Sp*Cas9:dsDNA (TTT)**	1	25	51 × 5	500	95.13 ± 0.13	-95.36 ± 0.16	95.19 ± 0.12	95.13 ± 0.07
	2		51 × 5	500	94.97 ± 0.13	-95.18 ± 0.15	95.06 ± 0.11	
dsDNA unbound *Sp*Cas9	1	25	51 × 10	1000	96.53 ± 0.06	-95.67 ± 0.06	96.19 ± 0.03	95.98 ± 0.10
	2		51 × 5	500	96.06 ± 0.07	-95.29 ± 0.08	95.86 ± 0.05	
	3		51 × 5	500	96.11 ± 0.09	-95.42 ± 0.09	95.76 ± 0.07	
	4	30	51 × 5	500	96.12 ± 0.10	-96.93 ± 0.07	96.13 ± 0.06	
	5	Full system	51 × 3	300	97.33 ± 0.15	-96.45 ± 0.14	96.74 ± 0.07	96.74 ± 0.07
***Sp*Cas9^{R1335A}: dsDNA (TGG)**	1	25	51 × 5	500	92.15 ± 0.11	-91.42 ± 0.11	91.88 ± 0.09	91.79 ± 0.09
	2		51 × 5	500	92.05 ± 0.12	-91.34 ± 0.11	91.70 ± 0.08	
***Sp*Cas9^{R1335A}: dsDNA (TGT)**	1	25	51 × 5	500	92.31 ± 0.18	-92.75 ± 0.16	92.14 ± 0.11	92.10 ± 0.05
	2		51 × 5	500	91.82 ± 0.18	-92.78 ± 0.16	92.05 ± 0.10	
***Sp*Cas9^{R1335A}: dsDNA (GAT)**	1	25	51 × 5	500	93.00 ± 0.11	-92.67 ± 0.16	92.71 ± 0.06	92.76 ± 0.05
	2		51 × 5	500	93.46 ± 0.11	-92.54 ± 0.12	92.82 ± 0.07	
***Sp*Cas9^{R1335A}: dsDNA (TTG)**	1	25	51 × 5	500	92.98 ± 0.13	92.83 ± 0.16	92.99 ± 0.12	92.83 ± 0.16
	2		51 × 5	500	92.88 ± 0.12	-92.55 ± 0.10	92.67 ± 0.11	
***Sp*Cas9^{R1335A}: dsDNA (TTT)**	1	25	51 × 5	500	94.08 ± 0.15	-93.29 ± 0.12	93.69 ± 0.10	93.62 ± 0.08
	2		51 × 5	500	93.73 ± 0.14	-93.44 ± 0.16	93.55 ± 0.12	
dsDNA unbound *Sp*Cas9^{R1335A}	1	25	51 × 5	500	95.50 ± 0.06	-95.12 ± 0.08	95.38 ± 0.05	95.32 ± 0.05
	2		51 × 5	500	95.38 ± 0.07	-95.29 ± 0.08	95.27 ± 0.05	

Table A2.3. MD trajectory averaged key interatomic distances (in Å) related to *Sp*Cas9 and canonical PAM (TGG) interactions in the *Sp*Cas9 and *Sp*Cas9^{E1219V} complex. The mean value of the distances obtained throughout the trajectories has been reported, while standard deviations are reported after '±'.

R1333, R1335:TGG PAM Interacting partners	**Trials**	***Sp*Cas9 (TGG) PAM**	***Sp*Cas9^{E1219V} (TGG) PAM**

R1333-NH1:G2-N7	Trial 1 (25 Å)	2.94 ± 0.12	2.98 ± 0.12
	Trial 2 (25 Å)	2.95 ± 0.23	2.96 ± 0.11
	Trial 3 (25 Å)	2.95 ± 0.10	2.97 ± 0.13
	Trial 4 (30 Å)	2.97 ± 0.13	2.95 ± 0.10
	Trial 5 (Full system)	2.96 ± 0.11	2.97 ± 0.14
	Average	**2.95 ± 0.14**	**2.96 ± 0.12**
R1333-NH2:G2-O6	Trial 1 (25 Å)	2.81 ± 0.15	2.86 ± 0.15
	Trial 2 (25 Å)	2.83 ± 0.20	2.85 ± 0.20
	Trial 3 (25 Å)	2.84 ± 0.14	2.87 ± 0.22
	Trial 4 (30 Å)	2.79 ± 0.11	2.82 ± 0.14
	Trial 5 (Full system)	2.85 ± 0.14	2.83 ± 0.15
	Average	**2.82 ± 0.15**	**2.85 ± 0.17**
R1333-NH1:G3-O6	Trial 1 (25 Å)	2.81 ± 0.11	2.77 ± 0.13
	Trial 2 (25 Å)	2.85 ± 0.15	2.83 ± 0.17
	Trial 3 (25 Å)	2.79 ± 0.10	3.05 ± 0.18
	Trial 4 (30 Å)	2.79 ± 0.12	2.83 ± 0.17
	Trial 5 (Full system)	2.93 ± 0.21	3.00 ± 0.12
	Average	**2.83 ± 0.14**	**2.90 ± 0.15**
R1333-NH2:G3-N7	Trial 1 (25 Å)	2.99 ± 0.11	3.01 ± 0.14
	Trial 2 (25 Å)	2.99 ± 0.11	2.99 ± 0.12
	Trial 3 (25 Å)	3.06 ± 0.16	3.00 ± 0.18
	Trial 4 (30 Å)	3.03 ± 0.11	3.04 ± 0.14
	Trial 5 (Full system)	3.03 ± 0.15	2.79 ± 0.14
	Average	**3.02 ± 0.13**	**2.97 ± 0.14**

Table A2.4. Trajectory averaged key interatomic distances (in Å) of *Sp*Cas9:TGT PAM interactions in the *Sp*Cas9 and *Sp*Cas9^{E1219V} complex. Values marked in red indicate no interaction.

R1333:TGT PAM Interacting partners	**Trials**	***SpCas9* (TGT) PAM**	***SpCas9*^{E1219V} (TGT) PAM**
R1333-NH1/2:G2-O6	Trial 1 (25 Å)	2.78 ± 0.34	2.99 ± 0.29
	Trial 2 (25 Å)	2.97 ± 0.20	2.81 ± 0.37
	Trial 3 (25 Å)	3.03 ± 0.42	3.32 ± 0.28
	Trial 4 (30 Å)	2.80 ± 0.14	2.79 ± 0.13
	Trial 5 (Full system)	2.93 ± 0.22	2.84 ± 0.16
	Average	**2.90 ± 0.26**	**2.95 ± 0.25**
	Trial 1 (25 Å)	2.99 ± 0.13	2.99 ± 0.14

R1333-NH1/2:G2-N7	Trial 2 (25 Å)	2.97 ± 0.12	2.92 ± 0.13
	Trial 3 (25 Å)	3.01 ± 0.16	3.38 ± 0.28
	Trial 4 (30 Å)	3.05 ± 0.14	2.98 ± 0.12
	Trial 5 (Full system)	3.02 ± 0.15	3.09 ± 0.16
	Average	**3.00 ± 0.14**	**3.07 ± 0.17**
R1333-NH1:T3-O4	Trial 1 (25 Å)	3.91 ± 0.26	3.29 ± 0.20
	Trial 2 (25 Å)	4.15 ± 0.34	2.78 ± 0.48
	Trial 3 (25 Å)	4.65 ± 0.51	3.02 ± 0.19
	Trial 4 (30 Å)	4.14 ± 0.59	3.14 ± 0.13
	Trial 5 (Full system)	4.97 ± 0.58	3.10 ± 0.18
	Average	**4.36 ± 0.46**	**3.07 ± 0.24**
R1333-NH2:T3-O4	Trial 1 (25 Å)	4.68 ± 0.43	4.77 ± 0.39
	Trial 2 (25 Å)	3.89 ± 0.49	4.00 ± 0.40
	Trial 3 (25 Å)	4.14 ± 0.35	4.65 ± 0.30
	Trial 4 (30 Å)	4.65 ± 0.30	4.05 ± 0.37
	Trial 5 (Full system)	5.68 ± 0.60	4.58 ± 0.25
	Average	**4.61 ± 0.43**	**4.41 ± 0.34**
R1333-NH1:G2-O2P	Trial 1 (25 Å)	5.25 ± 0.43	2.75 ± 0.12
	Trial 2 (25 Å)	3.74 ± 1.31	2.72 ± 0.09
	Trial 3 (25 Å)	5.10 ± 2.64	2.76 ± 0.13
	Trial 4 (30 Å)	10.21 ± 1.06	3.96 ± 0.32
	Trial 5 (Full system)	6.25 ± 0.46	3.78 ± 0.41
	Average	**6.11 ± 1.18**	**3.19 ± 0.21**
R1333-NH2:G2-O2P	Trial 1 (25 Å)	4.88 ± 0.38	2.92 ± 0.22
	Trial 2 (25 Å)	4.25 ± 0.86	3.12 ± 0.16
	Trial 3 (25 Å)	6.03 ± 2.61	2.93 ± 0.21
	Trial 4 (30 Å)	11.16 ± 1.30	3.18 ± 0.22
	Trial 5 (Full system)	8.54 ± 0.58	3.01 ± 0.23
	Average	**6.97 ± 1.15**	**3.03 ± 0.21**

Table A2.5. MD trajectory averaged key interatomic distances (in Å) related to *Sp*Cas9 and non-canonical GAT PAM interactions in the *Sp*Cas9 and *Sp*Cas9^{E1219V} complex.

R1333:GAT PAM Interacting partners	**Trials**	***SpCas9* (GAT) PAM**	***SpCas9*^{E1219V} (GAT) PAM**
	Trial 1 (25 Å)	5.34 ± 1.11	4.32 ± 0.45
	Trial 2 (25 Å)	4.56 ± 0.75	4.55 ± 0.73

R1333-NH1/2: A2-N6	Trial 3 (25 Å)	4.92 ± 1.34	4.96 ± 0.98
	Trial 4 (30 Å)	5.80 ± 1.35	4.62 ± 0.37
	Average	**5.16 ± 0.58**	**4.61 ± 0.45**
R1333-NH1/2: A2-N7	Trial 1 (25 Å)	5.59 ± 2.28	5.50 ± 0.45
	Trial 2 (25 Å)	6.02 ± 1.68	5.21 ± 1.42
	Trial 3 (25 Å)	4.71 ± 1.04	5.15 ± 1.08
	Trial 4 (30 Å)	6.17 ± 1.42	5.32 ± 0.59
	Average	**5.62 ± 0.83**	**5.29 ± 0.64**
R1333-NH1:T3-O4	Trial 1 (25 Å)	5.82 ± 1.75	6.25 ± 0.55
	Trial 2 (25 Å)	4.83 ± 1.22	4.04 ± 0.68
	Trial 3 (25 Å)	4.71 ± 0.94	4.76 ± 0.96
	Trial 4 (30 Å)	6.63 ± 0.95	5.23 ± 0.33
	Average	**5.50 ± 0.62**	**5.07 ± 0.45**
R1333-NH2:T3-O4	Trial 1 (25 Å)	5.95 ± 1.92	3.11 ± 0.29
	Trial 2 (25 Å)	5.43 ± 1.39	2.97 ± 0.30
	Trial 3 (25 Å)	4.40 ± 0.81	3.02 ± 0.21
	Trial 4 (30 Å)	5.88 ± 1.11	3.08 ± 0.21
	Average	**5.41 ± 0.68**	**3.05 ± 0.17**

Table A2.6. MD trajectory averaged key interatomic distances (in Å) related to *Sp*Cas9 and non-canonical TTG PAM interactions in the *Sp*Cas9 and *Sp*Cas9^{E1219V} complex.

R1335: TTG PAM Interacting partners	**Trials**	***Sp*Cas9:dsDNA (TTG) PAM**	***Sp*Cas9^{E1219V}:dsDNA (TTG) PAM**
R1335-NH1/2:G3-N7	Trial 1	3.02 ± 0.15	2.83 ± 0.14
	Trial 2	3.08 ± 0.18	3.08 ± 0.15
	Average	**3.05 ± 0.17**	**2.96 ± 0.15**
R1335-NH1/2:G3-O6	Trial 1	2.80 ± 0.12	2.99 ± 0.14
	Trial 2	2.79 ± 0.13	2.78 ± 0.12
	Average	**2.79 ± 0.13**	**2.88 ± 0.13**

Table A2.7. MD trajectory averaged key interatomic distances (in Å) related to *Sp*Cas9 and non-canonical TTT PAM interactions in the *Sp*Cas9 and *Sp*Cas9^{E1219V} complex.

R1333: TTG PAM Interacting partners	**Trials**	***Sp*Cas9:dsDNA (TTT) PAM**	***Sp*Cas9^{E1219V}:dsDNA (TTT) PAM**
	Trial 1	5.85 ± 0.93	3.27 ± 0.24

R1333-NH1/2:T2-O4	Trial 2	5.82 ± 0.91	3.24 ± 0.33
	Average	**5.83 ± 0.92**	**3.26 ± 0.29**
R1335-NH1/2:T3-O4	Trial 1	7.93 ± 1.22	8.62 ± 1.16
	Trial 2	5.26 ± 0.39	3.38 ± 0.60
	Average	**6.59 ± 0.80**	**6.00 ± 0.88**

Table A2.8. Selected interatomic distances (in Å) obtained from the MD trajectories of *Sp*Cas9^{R1335A} (single-mutant *Sp*Cas9) and SpCas9$^{R1335A,\ E1219V}$ (double-mutant *Sp*Cas9) complexed with different PAM sequences.

R1333: PAM Interacting partners	**Trials**	**System**	
		***Sp*Cas9^{R1335A}:dsDNA (TGG) PAM**	***Sp*Cas9$^{R1335A,\ E1219V}$: dsDNA (TGG) PAM**
R1333-NH1/2:G2-N7	Trial 1	4.65 ± 1.10	2.93 ± 0.10
	Trial 2	4.00 ± 0.63	2.96 ± 0.11
	Average	**4.33 ± 0.87**	**2.94 ± 0.10**
R1333-NH1/2:G2-O6	Trial 1	4.97 ± 0.79	3.79 ± 0.32
	Trial 2	3.93 ± 0.66	3.38 ± 0.21
	Average	**4.45 ± 0.73**	**3.58 ± 0.26**
R1333-NH1/2:G3-O6	Trial 1	5.19 ± 1.15	3.11 ± 0.15
	Trial 2	5.31 ± 0.52	3.24 ± 0.10
	Average	**5.25 ± 0.83**	**3.17 ± 0.12**
		***Sp*Cas9^{R1335A}:dsDNA (TGT) PAM**	***Sp*Cas9$^{R1335A,\ E1219V}$: dsDNA (TGT) PAM**
R1333-NH1:G2-N7	Trial 1	3.98 ± 0.75	2.98 ± 0.13
	Trial 2	4.95 ± 0.86	3.00 ± 0.17
	Average	**4.46 ± 0.80**	**2.99 ± 0.15**
R1333-NH1:G2-O6	Trial 1	4.52 ± 0.72	3.15 ± 0.15
	Trial 2	4.66 ± 0.83	2.94 ± 0.34
	Average	**4.59 ± 0.77**	**3.07 ± 0.24**
R1333-NH2:T3-O4	Trial 1	4.51 ± 0.61	3.05 ± 0.16
	Trial 2	4.99 ± 0.78	3.07 ± 0.28
	Average	**4.75 ± 0.69**	**3.06 ± 0.22**
		***Sp*Cas9^{R1335A}:dsDNA (GAT) PAM**	***Sp*Cas9$^{R1335A,\ E1219V}$: dsDNA (GAT) PAM**
	Trial 1	6.07 ± 0.93	6.17 ± 0.44
	Trial 2	6.84 ± 1.27	5.88 ± 0.72

R1333-NH1:A2-N6	**Average**	**6.45 ± 1.13**	**6.02 ± 0.58**
R1333-NH2:A2-N7	Trial 1	5.73 ± 0.92	5.43 ± 0.42
	Trial 2	5.70 ± 1.32	5.02 ± 0.91
	Average	**5.72 ± 1.12**	**5.22 ± 0.67**
R1333-NH1:T3-O4	Trial 1	6.13 ± 1.06	2.98 ± 0.26
	Trial 2	5.51 ± 1.04	3.22 ± 0.15
	Average	**5.82 ± 1.05**	**3.10 ± 0.21**
		***Sp*Cas9^{R1335A}:dsDNA(TTG) PAM**	***Sp*Cas9$^{R1335A, E1219V}$: dsDNA(TTG) PAM**
R1333-NH1/2:G3-N7	Trial 1	7.87 ± 1.01	3.15 ± 0.17
	Trial 2	6.88 ± 0.73	3.62 ± 0.29
	Average	**7.37 ± 0.85**	**3.39 ± 0.24**
R1333-NH1/2:G3-O6	Trial 1	5.73 ± 0.93	3.07± 0.27
	Trial 2	5.39 ± 1.02	3.09 ± 0.18
	Average	**5.56 ± 0.98**	**3.08 ± 0.22**
		***Sp*Cas9^{R1335A}:dsDNA(TTT) PAM**	***Sp*Cas9$^{R1335A, E1219V}$: dsDNA(TTT) PAM**
R1333-NH1/2:T2-O4	Trial 1	4.88 ± 0.96	3.19 ± 0.52
	Trial 2	5.90 ± 1.56	3.14 ± 0.25
	Average	**5.39 ± 1.26**	**3.16 ± 0.38**

Table A2.9. Trajectory averaged number of water molecules within 4 Å of key PAM interacting atoms in the wild-type, single-mutant, and double-mutant *Sp*Cas9 in complex with canonical and non-canonical PAM sequences. Standard deviation values were reported after '±' symbol.

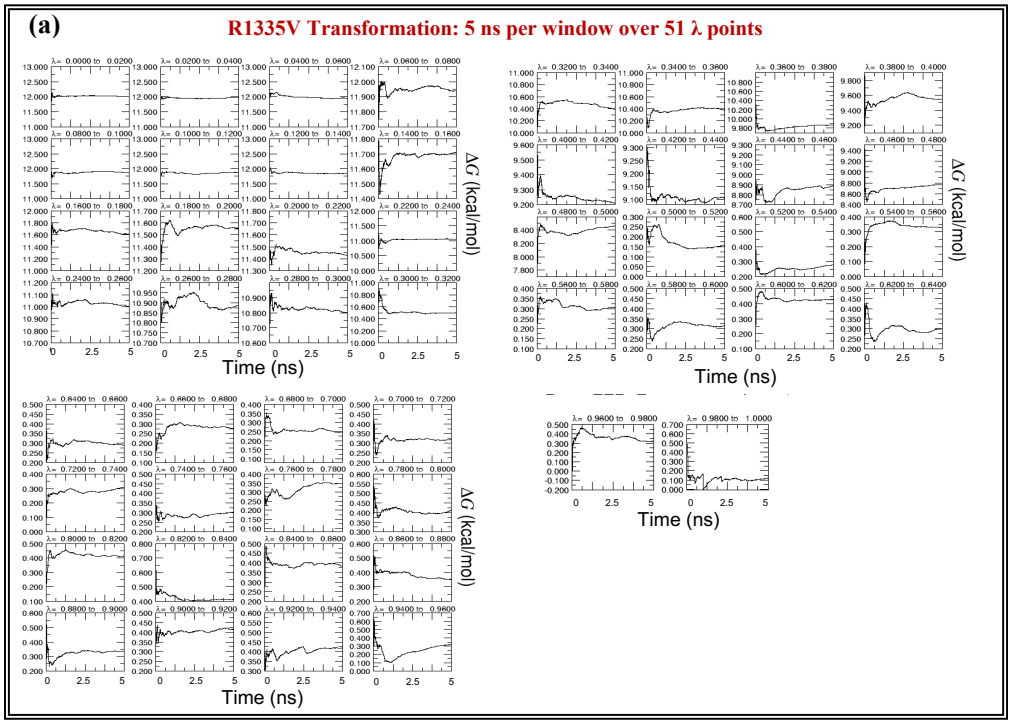
Structure	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5 (Full system)	Average
Water molecules within 4 Å of R1333 (NH1 and NH2), R1335 (NH1 and NH2), G2 (N7 and O6) and G3 (N7 and O6)						
***Sp*Cas9:dsDNA (TGG)**	19.48 ± 2.80	19.64 ± 2.48	19.2 ± 2.73	19.76 ± 3.30	19.56 ± 3.36	**19.53 ± 2.93**
***Sp*Cas9^{E1219V}:dsDNA (TGG)**	19.82 ± 2.52	19.68 ± 2.47	19.44 ± 2.75	19.16 ± 2.44	19.96 ± 2.84	**19.61 ± 2.60**
Water molecules within 4 Å of R1333 (NH1 and NH2), R1335 (NH1 and NH2), G2 (N7 and O6) and T3 (C5 and O4)						
***Sp*Cas9:dsDNA (TGT)**	21.62 ± 3.06	20.74 ± 3.46	20.82 ± 3.35	20.64 ± 3.38	21.42 ± 2.87	**21.05 ± 3.22**
***Sp*Cas9^{E1219V}:dsDNA (TGT)**	13.2 ± 3.01	11.64 ± 2.47	11.18 ± 2.32	11.66 ± 2.40	11.98 ± 2.48	**11.93 ± 2.54**
Water molecules within 4 Å of R1333 (NH1 and NH2), A2 (N6 and N7) and T3 (C5 and O4)						
***Sp*Cas9:dsDNA (GAT)**	23.08 ± 4.20	25.94 ± 3.50	20.27 ± 4.32	24.14 ± 4.21	-	**23.35 ± 4.06**

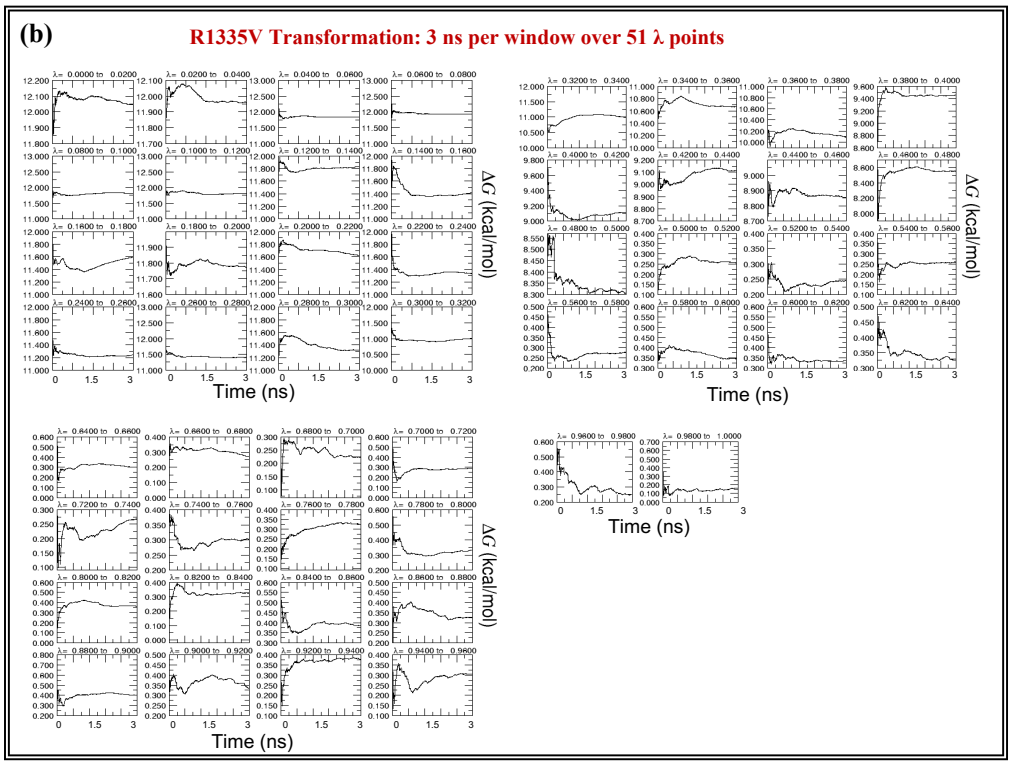
***Sp*Cas9^{E1219V}:dsDNA (GAT)**	15.41 ± 3.34	16.06 ± 3.38	15.22 ± 3.04	15.20 ± 3.44	-	**15.47 ± 3.30**
Water molecules within 4 Å of R1335 (NH1 and NH2), T2 (C5 and O4) and G2 (N7 and O6)						
***Sp*Cas9:dsDNA (TTG)**	15.82 ± 3.02	14.88 ± 3.21	-	-	-	**15.35 ± 3.12**
***Sp*Cas9^{E1219V}:dsDNA (TTG)**	15.14 ± 3.17	15.38 ± 2.57	-	-	-	**15.26 ± 2.87**
Water molecules within 4 Å of R1333 (NH1 and NH2), R1335 (NH1 and NH2), T2 (C5 and O4) and T3 (C5 and O4)						
***Sp*Cas9:dsDNA (TTT)**	26.04 ± 3.34	28.44 ± 3.39	-	-	-	**27.24 ± 3.37**
***Sp*Cas9^{E1219V}:dsDNA (TTT)**	19.28 ± 3.19	15.18 ± 3.37	-	-	-	**17.33 ± 3.28**
Water molecules within 4 Å of R1333 (NH1 and NH2), G2 (N7 and O6) and G3 (N7 and O6)						
***Sp*Cas9^{R1335A}:dsDNA (TGG)**	23.78 ± 3.48	21.22 ± 3.68	-	-	-	**22.50 ± 3.58**
***Sp*Cas9^{R1335A,E1219V}:dsDNA (TGG)**	12.62 ± 1.61	11.40 ± 1.33	-	-	-	**12.01 ± 1.47**
Water molecules within 4 Å of R1333 (NH1 and NH2), G2 (N7 and O6) and T3 (C5 and O4)						
***Sp*Cas9^{R1335A}:dsDNA (TGT)**	20.8 ± 5.25	19.86 ± 3.68	-	-	-	**20.33 ± 4.46**
***Sp*Cas9^{R1335A,E1219V}:dsDNA (TGT)**	12.1 ± 2.33	12.42 ± 2.56	-	-	-	**12.26 ± 2.45**
Water molecules within 4 Å of R1333 (NH1 and NH2), A2 (N6 and N7) and T3 (C5 and O4)						
***Sp*Cas9^{R1335A}:dsDNA (GAT)**	25.9 ± 5.78	29.04 ± 5.75	-	-	-	**27.47 ± 5.77**
***Sp*Cas9^{R1335A,E1219V}:dsDNA (GAT)**	18.36 ± 3.08	17.78 ± 3.36	-	-	-	**18.07 ± 3.22**
Water molecules within 4 Å of R1333 (NH1 and NH2), T2 (C5 and O4) and G3 (N7 and O6)						
***Sp*Cas9^{R1335A}:dsDNA (TTG)**	24.9 ± 4.92	22.7 ± 3.7	-	-	-	**23.80 ± 4.31**
***Sp*Cas9^{R1335A,E1219V}:dsDNA (TTG)**	17.38 ± 3.09	17.26 ± 2.88	-	-	-	**17.32 ± 2.99**
Water molecules within 4 Å of R1333 (NH1 and NH2), T2 (C5 and O4) and T3 (C5 and O4)						
***Sp*Cas9^{R1335A}:dsDNA (TTT)**	17.16 ± 4.10	20.1 ± 2.91	-	-	-	**18.63 ± 3.5**
***Sp*Cas9^{R1335A,E1219V}:dsDNA (TTT)**	11.36 ± 2.38	10.18 ± 2.58	-	-	-	**10.77 ± 2.48**

Table A2.10. Trajectory averaged number of water molecules within 4 Å of C5 of Thy 3 in the wild-type, single-mutant, and double-mutant *Sp*Cas9 in complex with various PAM sequences.

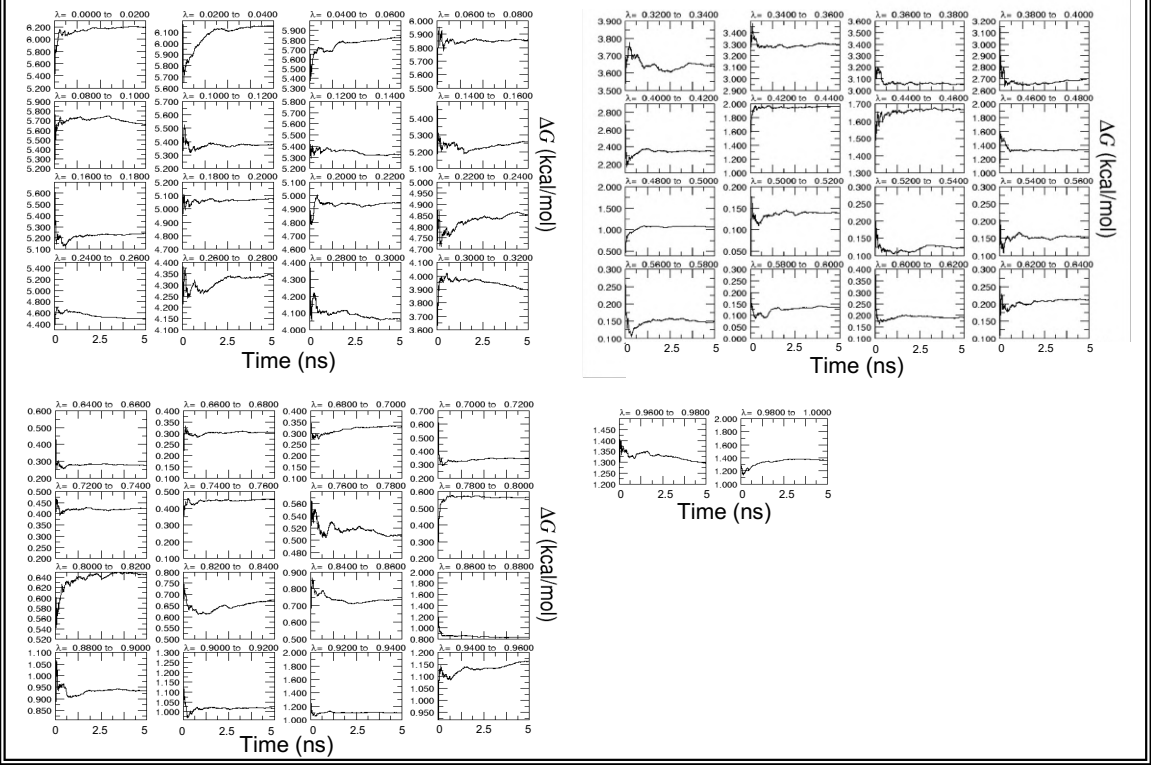
Structure	Water molecules within 4 Å of C5 of Thy 3					
	Trial 1	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5 (Full system)**	**Average**
*Sp*Cas9:dsDNA (TGT)	2.87 ± 1.28	2.78 ± 1.28	2.56 ± 1.03	2.88 ± 1.08	2.48 ± 0.84	2.71 ± 1.10
*Sp*Cas9^{E1219V}:dsDNA (TGT)	0.86 ± 0.79	0.66 ± 0.66	0.42 ± 0.61	0.68 ± 0.65	0.78 ± 0.79	0.68 ± 0.70
*Sp*Cas9^{R1335A}:dsDNA (TGT)	2.4 ± 1.02	2.48 ± 1.46	-	-	-	2.44 ± 1.24
*Sp*Cas9$^{R1335A, E1219V}$:dsDNA (TGT)	0.74 ± 0.69	0.54 ± 0.70	-	-	-	0.64 ± 0.69
*Sp*Cas9:dsDNA (GAT)	2.37 ± 1.10	2.40 ± 1.18	2.08 ± 0.88	2.54 ± 0.97	-	2.35 ± 1.03
*Sp*Cas9^{E1219V}:dsDNA (GAT)	0.79 ± 0.75	0.64 ± 0.68	0.84 ± 0.71	0.62 ± 0.64	-	0.72 ± 0.69
*Sp*Cas9^{R1335A}:dsDNA (GAT)	2.94 ± 1.35	2.35 ± 1.35	-	-	-	2.64 ± 1.35
*Sp*Cas9$^{R1335A, E1219V}$:dsDNA (GAT)	0.6 ± 0.66	0.68 ± 0.61	-	-	-	0.64 ± 0.63
*Sp*Cas9:dsDNA (TTT)	2.42 ± 1.98	1.54 ± 1.47	-	-	-	1.98 ± 1.72
*Sp*Cas9^{E1219V}:dsDNA (TTT)	0.74 ± 0.66	0.66 ± 0.62	-	-	-	0.70 ± 0.64
*Sp*Cas9^{R1335A}:dsDNA (TTT)	1.98 ± 1.11	1.74 ± 1.43	-	-	-	1.86 ± 1.26
*Sp*Cas9$^{R1335A, E1219V}$:dsDNA (TTT)	0.64 ± 0.59	0.56 ± 0.61	-	-	-	0.60 ± 0.60

Chapter 3

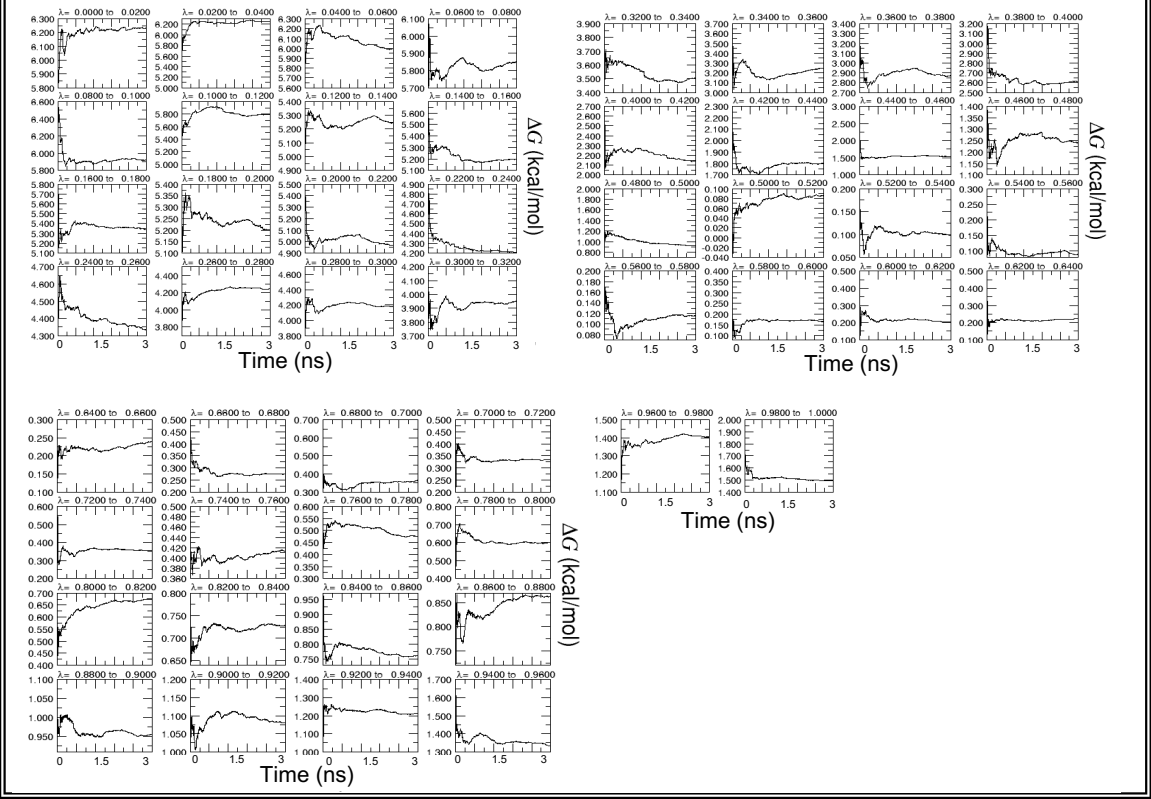




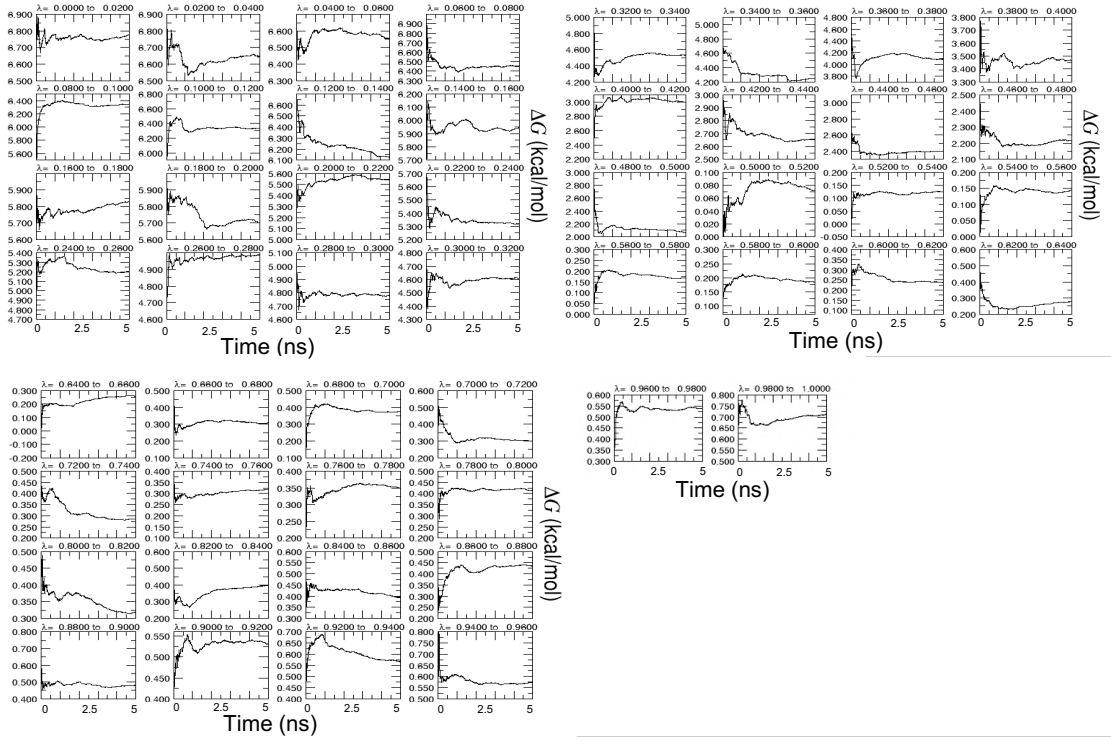
(c) **E1219F Transformation: 5 ns per window over 51 λ points**



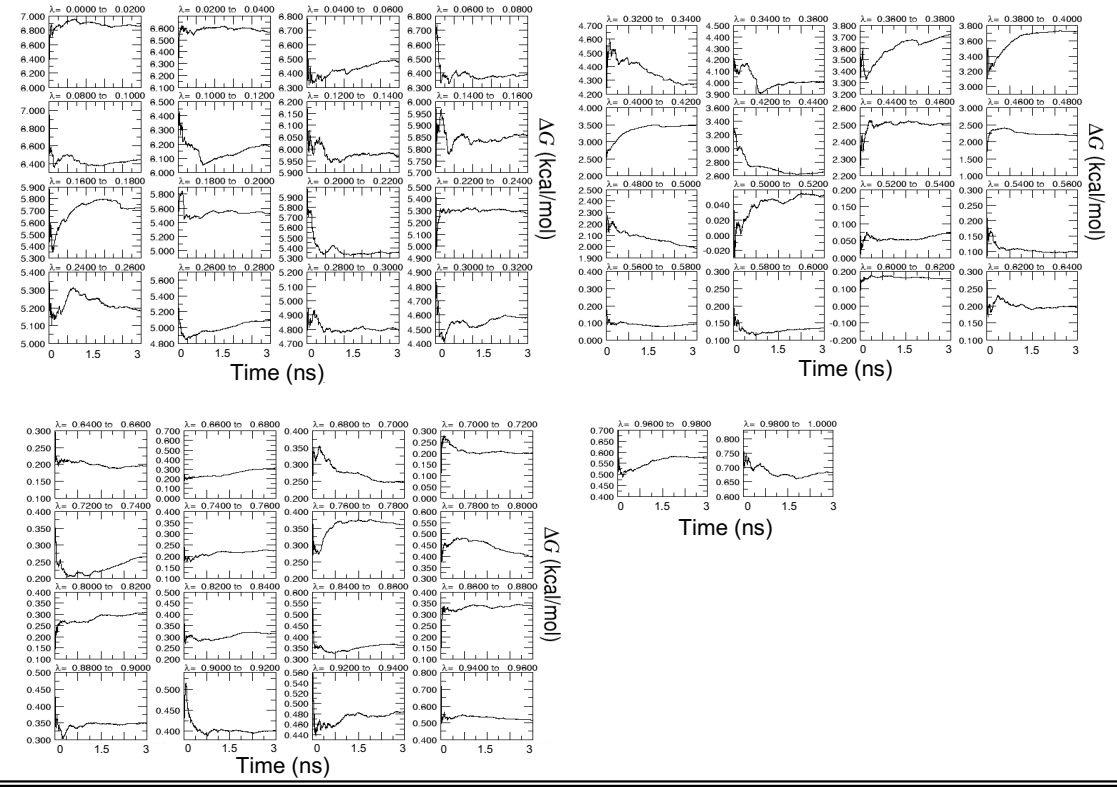
(d) **E1219F Transformation: 3 ns per window over 51 λ points**



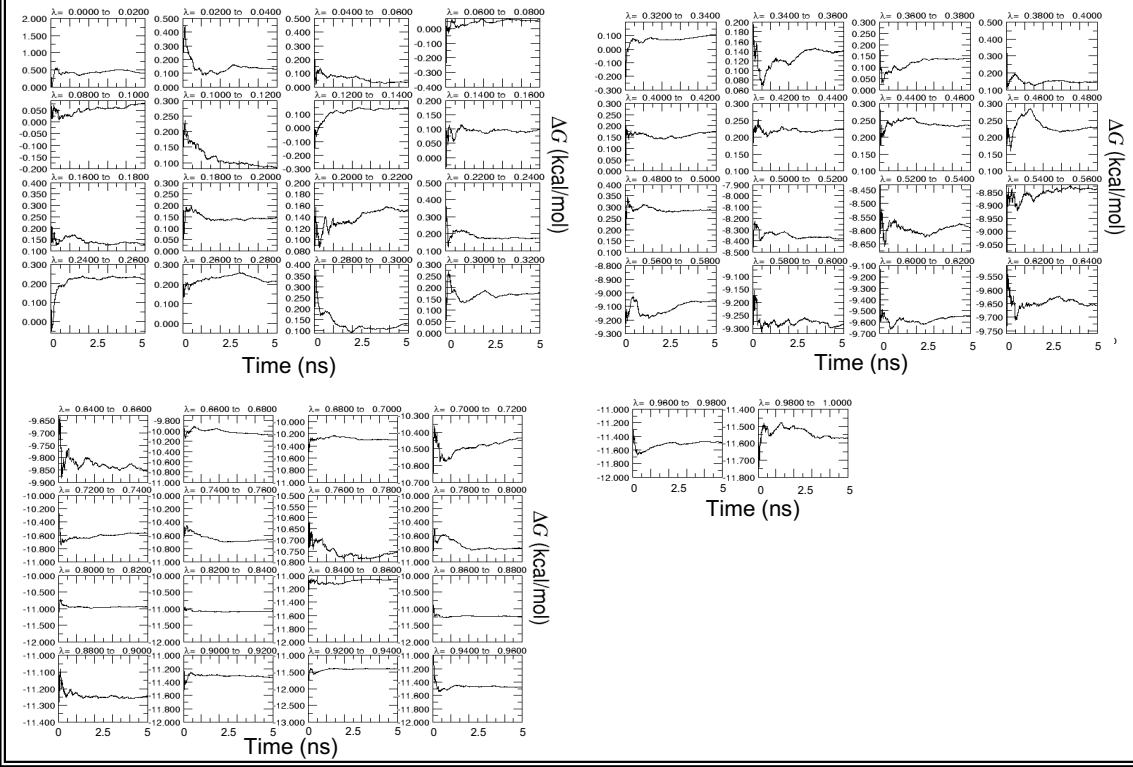
(e) **D1135V Transformation: 5 ns per window over 51 λ points**



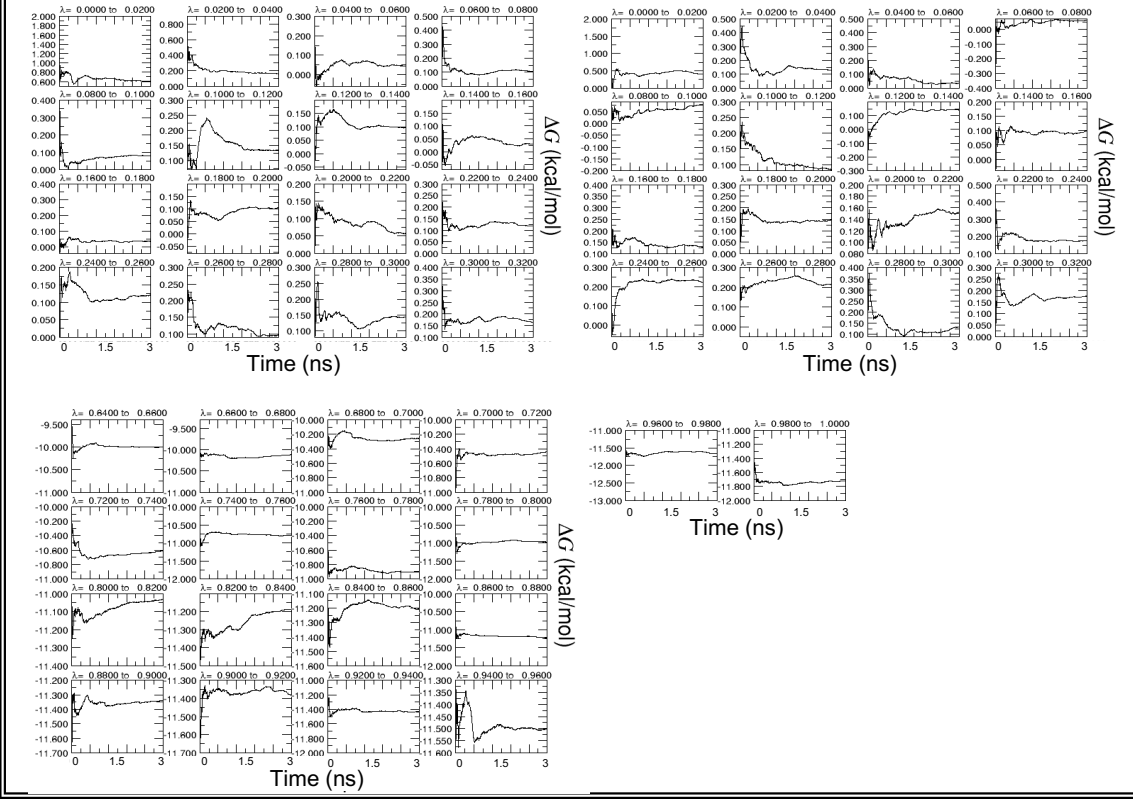
(f) **D1135V Transformation: 3 ns per window over 51 λ points**



(g) **L1111R Transformation: 5 ns per window over 51 λ points**



(h) **L1111R Transformation: 3 ns per window over 51 λ points**



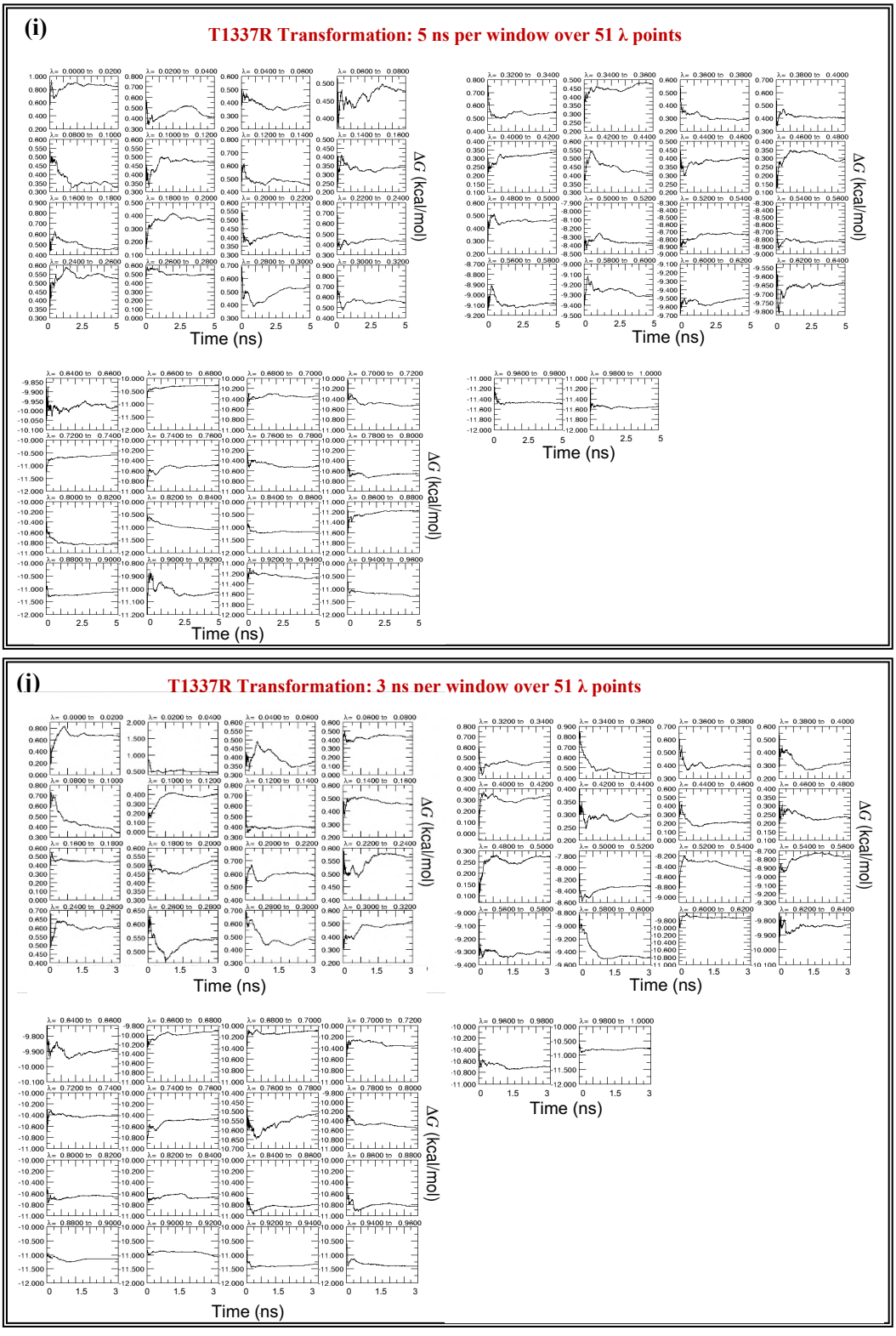


Figure A3.1. Time-dependent free energy plots (ΔG vs. time for each neighboring λ window) for alchemical transformation performed at either 5 ns or 3 ns per window over 51 λ points.

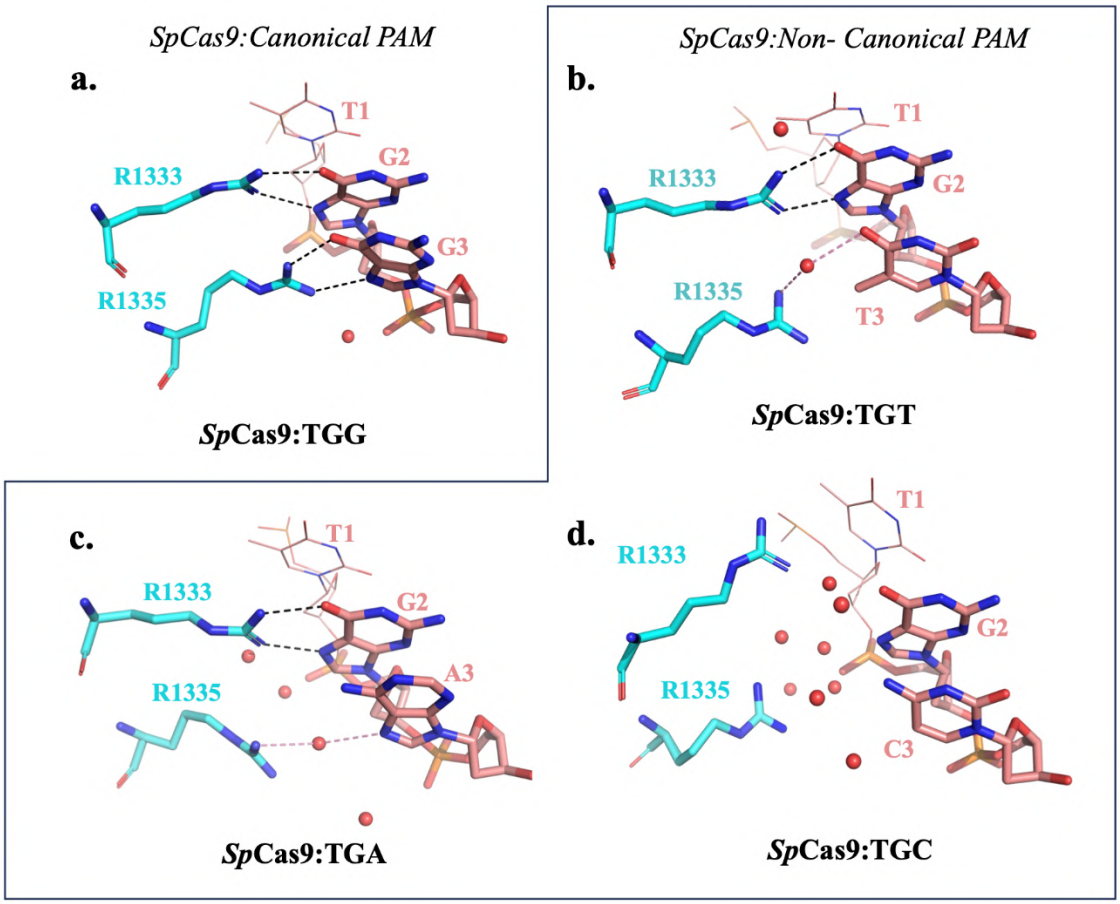


Figure A3.2. Comparison of the *Sp*Cas9:PAM interaction network of *Sp*Cas9 bound to (a) canonical TGG PAM and (b, c, d) Non-canonical TGT, TGT, TGC PAM. Water molecules are represented as red spheres. Black dotted lines represent direct *Sp*Cas9:PAM interactions, while pink dotted lines represent water-mediated interactions.

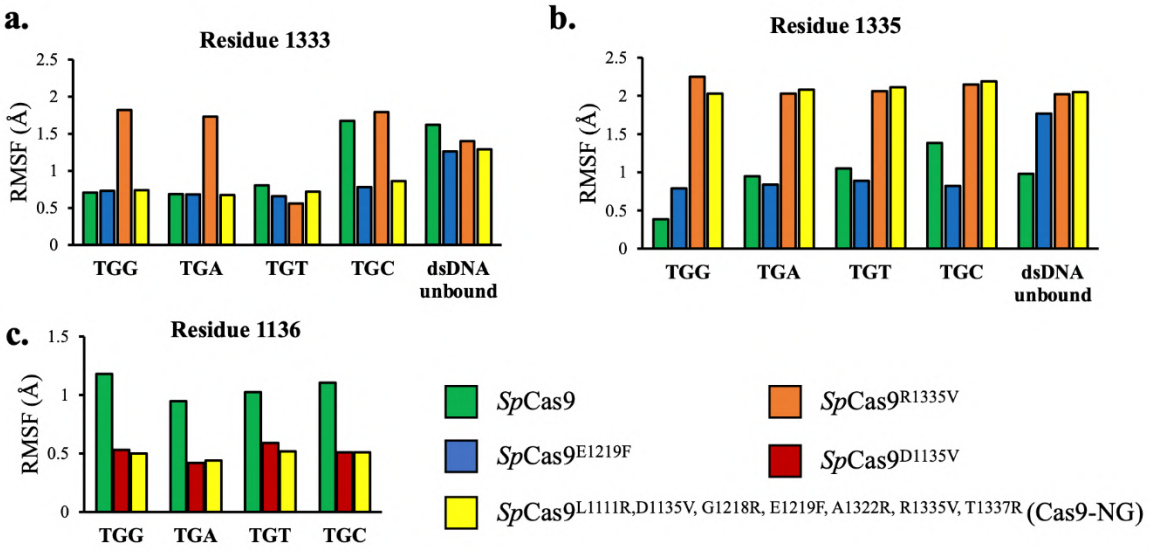


Figure A3.3. Root Mean Square Fluctuation (RMSF) analysis of residues (a)1333, (b) R1335, (c) 1136 in wild-type *Sp*Cas9 and its mutants bound to different PAM sequences (TGG, TGA, TGT, TGC) and unbound dsDNA conditions.

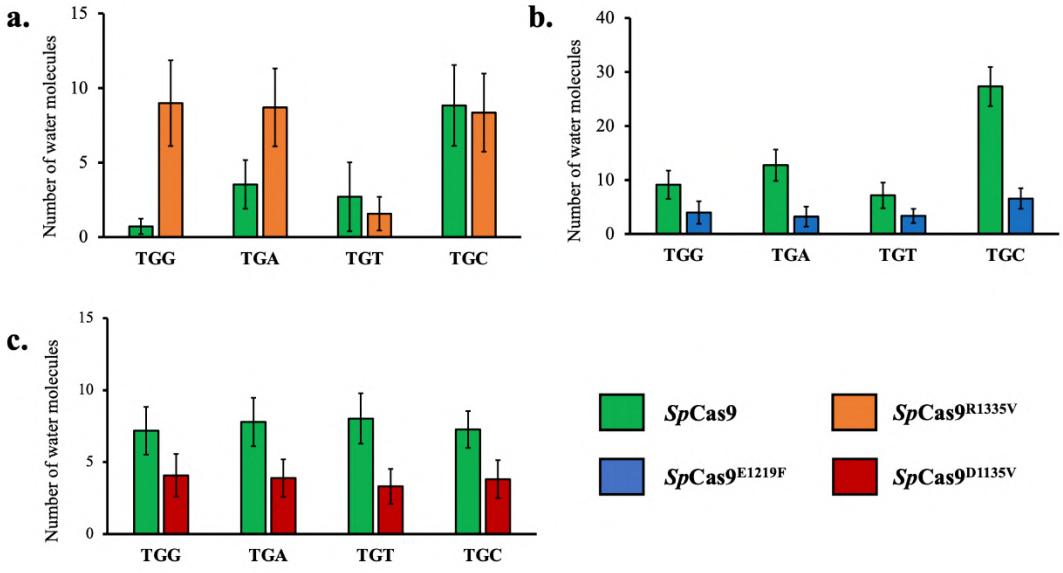


Figure A3.4. Trajectory averaged (each over three trials) number of water molecules within 4 Å of (a) Watson crick edge of second and third PAM nucleotides of *Sp*Cas9 and *Sp*Cas9^{R1335V}, (b) PAM interacting atoms (NH1, NH2 atoms of R1333, R1335 and Watson crick edge of second and third PAM nucleotides) for *Sp*Cas9 and *Sp*Cas9^{E1219F} and (c) OG atom of S1136 residue. The values are shown for SpCas9 and SpCas9^{Mutant} bound to four different PAM sequences: TGG, TGA, TGT, TGC

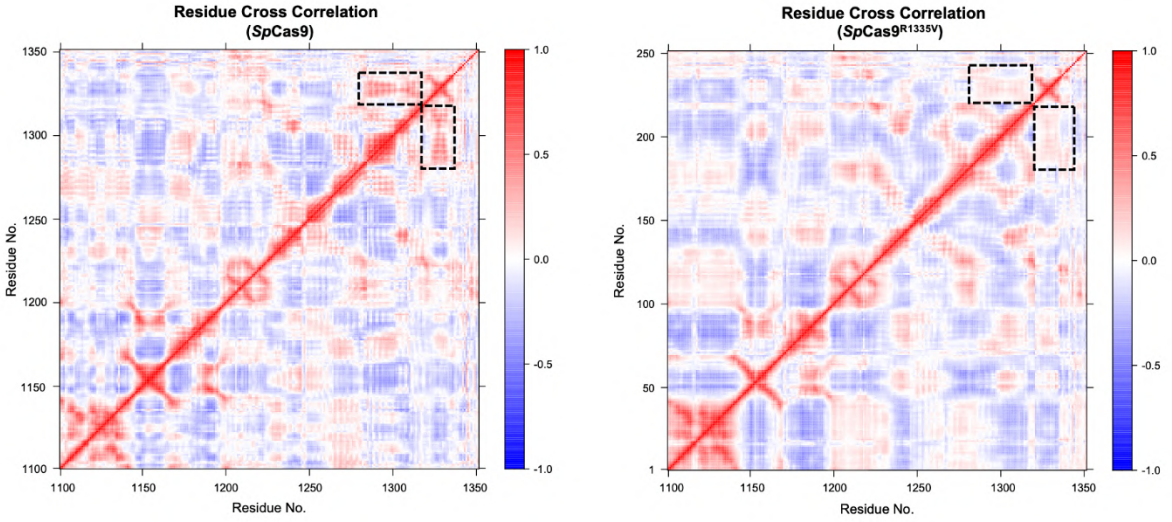


Figure A3.5. DCCM for residues of PI domain (1100-1350) of *Sp*Cas9 (left) and the *Sp*Cas9^{R1335V} mutant (right). The colour scale ranges from –1.0 (blue, anti-correlated) to +1.0 (red, positively correlated). The diagonal (red) represents self-correlation. Off-diagonal differences (shown in boxes) highlighted the loss of correlation in the PAM binding pocket in response to R1335V mutation (left →right).

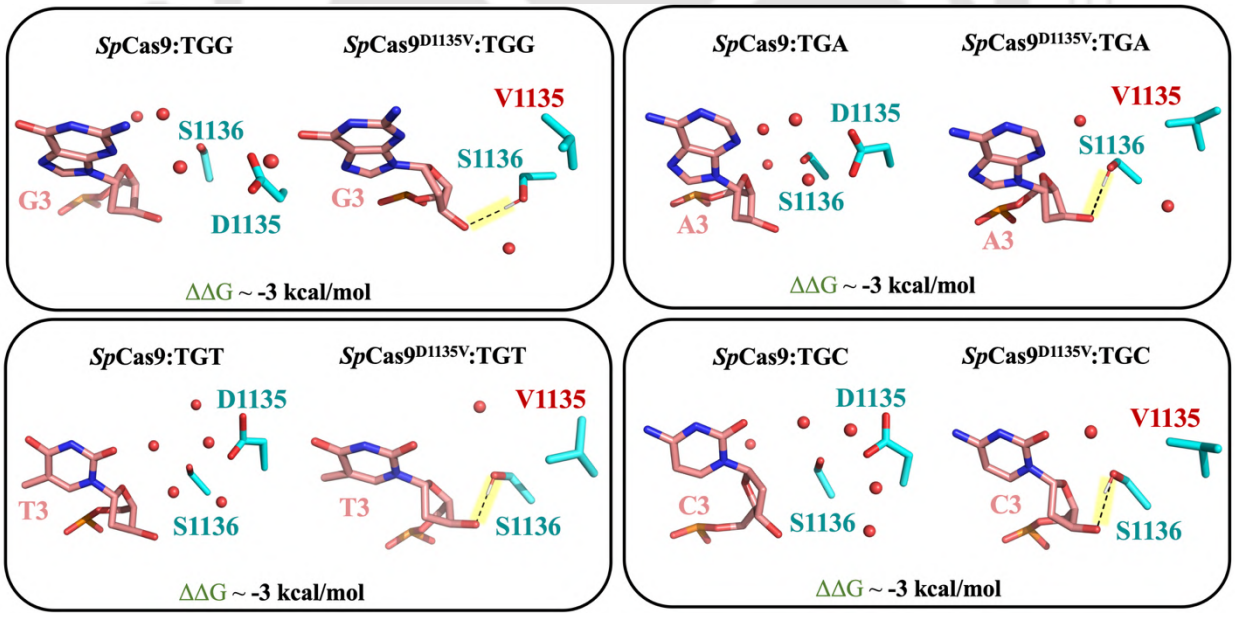


Figure A3.6. Comparison of the structures of precatalytic *Sp*Cas9 and *Sp*Cas9^{D1135V} bound to different PAM sequences. Water molecules are represented as red spheres.

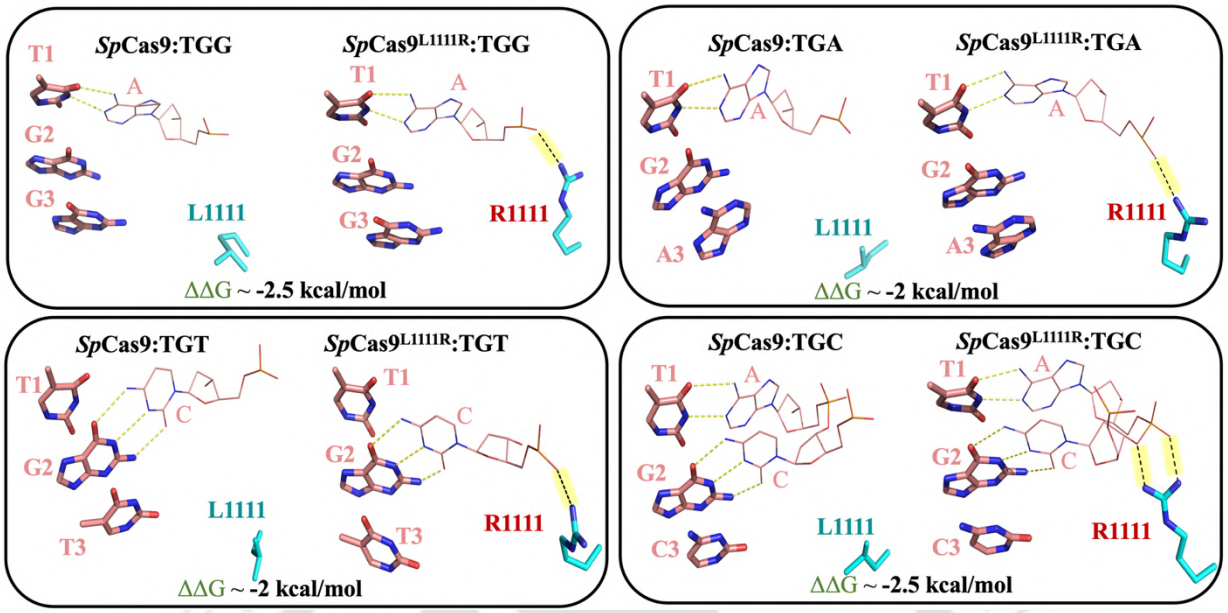


Figure A3.7. Comparison of the structures of precatalytic *Sp*Cas9 and *Sp*Cas9^{L1111R} bound to different PAM sequences. Water molecules are represented as red spheres.

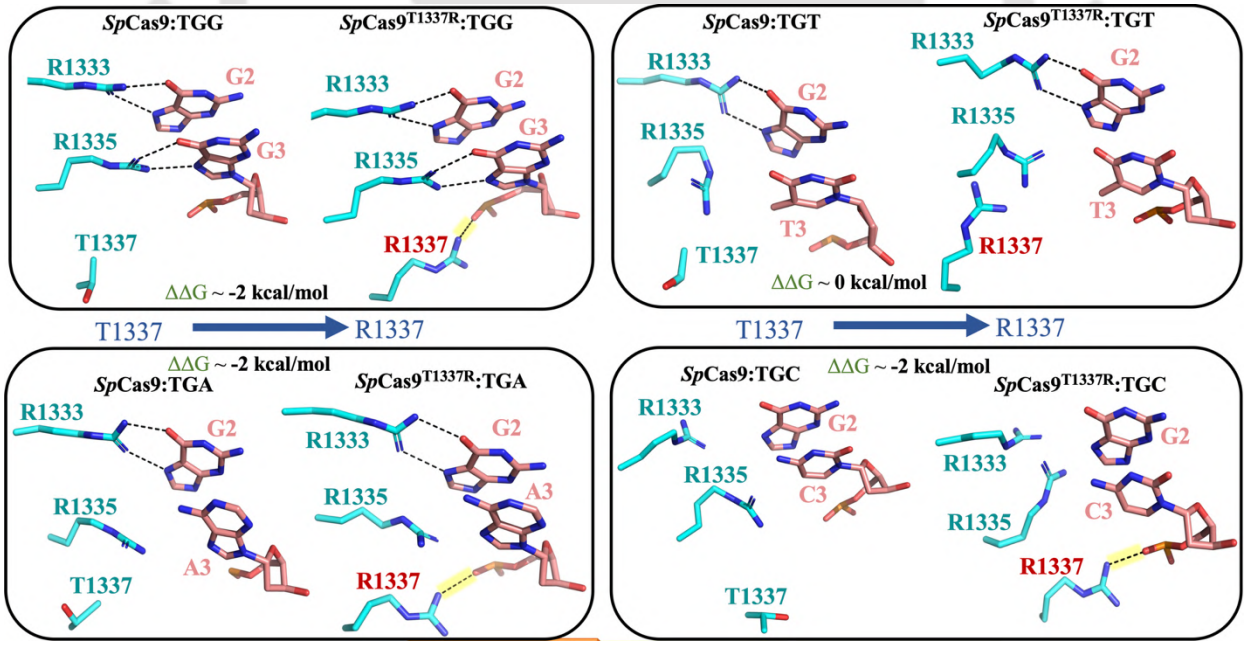


Figure A3.8. Comparison of the structures of precatalytic *Sp*Cas9 and *Sp*Cas9^{T1337R} bound to different PAM sequences. Water molecules are represented as red spheres.

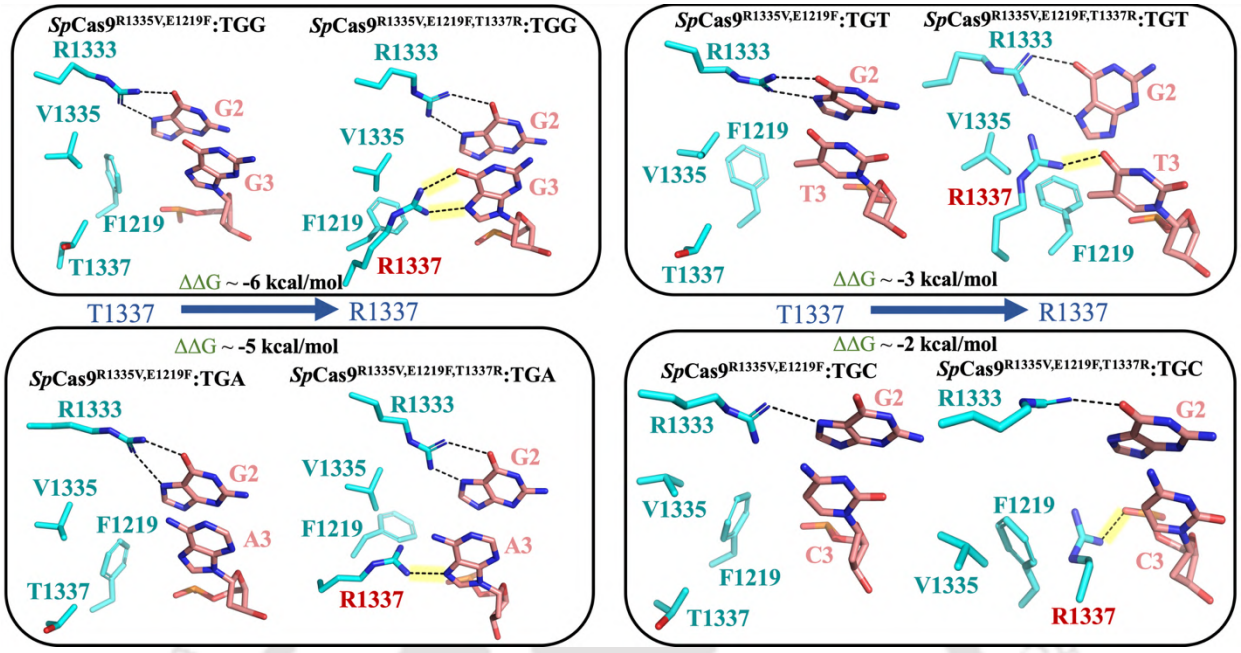


Figure A3.9. Comparison of the structures of precatalytic *Sp*Cas9^{R1335V,E1219F} and *Sp*Cas9R$^{1335V, E1219F, T1337R}$ bound to different PAM sequences. Water molecules are represented as red spheres.

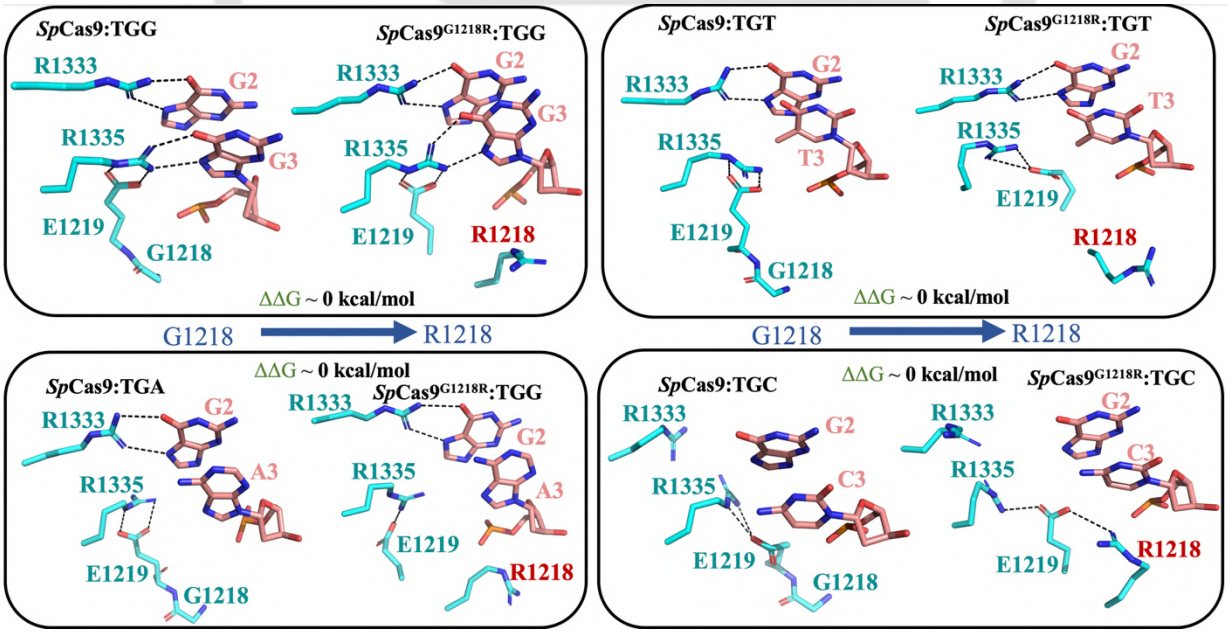


Figure A3.10. Comparison of the structures of precatalytic *Sp*Cas9 and *Sp*Cas9^{G1218R} bound to different PAM sequences.

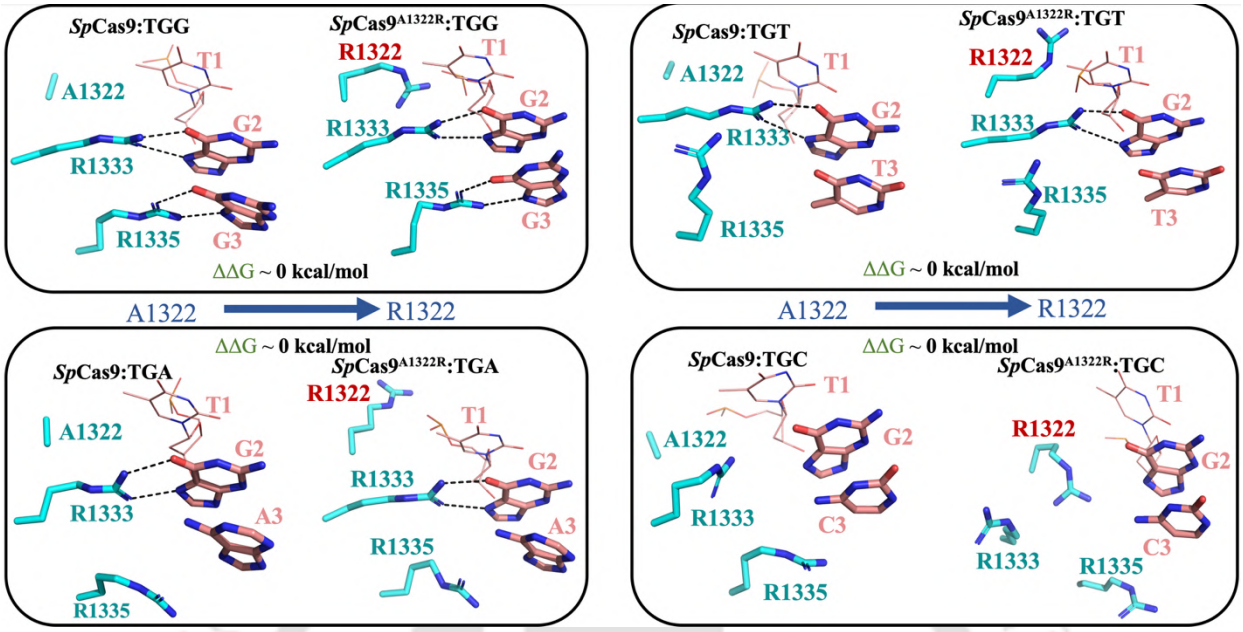


Figure A3.11. Comparison of the structures of precatalytic *Sp*Cas9 and *Sp*Cas9^{A1322R} bound to different PAM sequences.

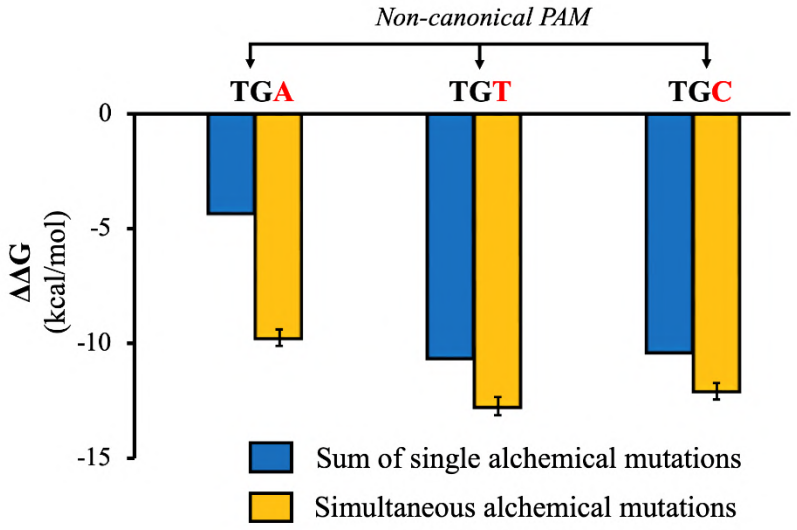


Figure A3.12. Comparison of the values of ΔΔG was determined by transforming seven residues simultaneously (yellow), in contrast to the data obtained by summing the energetic contributions from the transformations of individual amino acids (blue) for *Sp*Cas9 bound to non-cognate PAM (TGA, TGT, TGC).

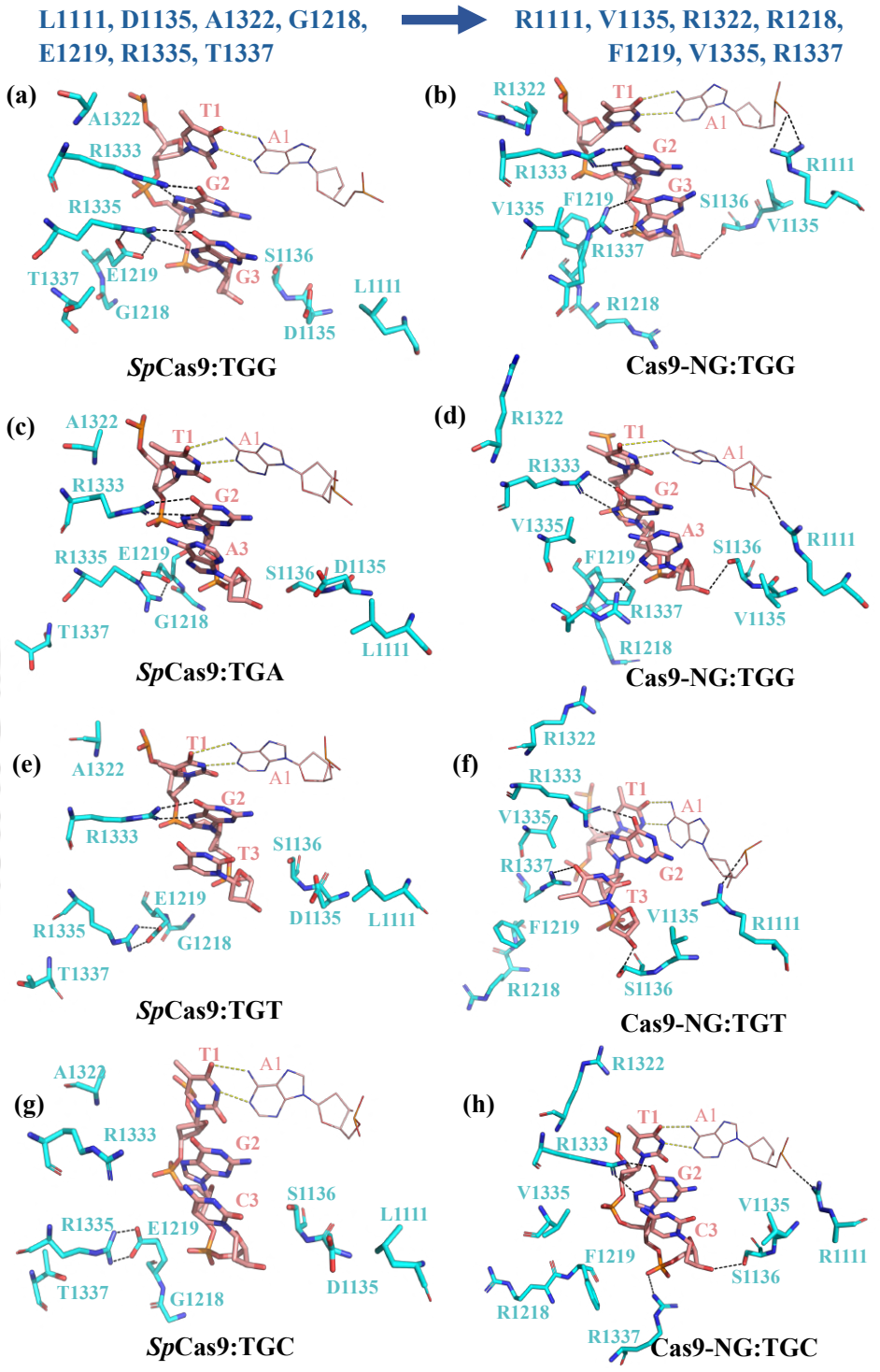


Figure A3.13. Comparison of the structures of precatalytic *Sp*Cas9 (left: a, c, e, g) and computationally simulated Cas9-NG (*Sp*Cas9$^{R1335V, E1219F, D1135V, L1111R, T1337R, G1218R, A1322R}$) PAM binding region (left: b, d, f, h) bound to different PAM sequences: (a, b) 5`-TGG-3` , (c, d) 5`-TGA-3` , (e, f) 5`-TGT-3`, (g, h) 5`-TGC-3`.

Table A3.1. List of different simulation models used in this study, their size, box dimension, solvation, and counter ions used.

Simulation Model	Spherically truncated around residue	Box Dimensions (Å^3)	Total number of atoms	Number of waters	Number of Na+	Number of Cl-	Number of trials
*Sp*Cas9:dsDNA (TGG)	1335	85.2 × 79.2 × 77.7	48218	14417	37	0	3
*Sp*Cas9:dsDNA (TGG)	Full system	133.2 × 123.7 × 136.7	210509	60886	133	0	1
*Sp*Cas9:dsDNA (TGA)	1335	85.2 × 79.2 × 77.7	48114	14382	37	0	3
*Sp*Cas9:dsDNA (TGT)	1335	85.2 × 79.2 × 77.7	48228	14420	37	0	3
*Sp*Cas9:dsDNA (TGC)	1335	85.2 × 79.2 × 77.7	48311	14448	37	0	3
dsDNA unbound *Sp*Cas9	1335	85.2 × 79.2 × 69.6	43301	13074	13	0	3
dsDNA unbound *Sp*Cas9	Full system	141.7 × 115.7 × 123.3	188387	54511	43	0	1
*Sp*Cas9:dsDNA (TGG)	1219	78.6 × 82.1 × 77.7	45876	13480	28	0	3
*Sp*Cas9:dsDNA (TGA)	1219	78.6 × 82.1 × 77.7	45883	13482	28	0	3
*Sp*Cas9:dsDNA (TGT)	1219	78.6 × 82.1 × 77.7	45868	13477	28	0	3
*Sp*Cas9:dsDNA (TGC)	1219	78.6 × 82.1 × 77.7	45885	13483	28	0	3
dsDNA unbound *Sp*Cas9	1219	82.2 × 81.5 × 73.6	45591	13780	6	0	3
*Sp*Cas9:dsDNA (TGG)	1135	85.7 × 81.4 × 77.3	49504	14938	40	0	3
*Sp*Cas9:dsDNA (TGA)	1135	85.7 × 81.4 × 77.3	49535	14948	40	0	3
*Sp*Cas9:dsDNA (TGT)	1135	85.7 × 81.4 × 77.3	49511	14940	40	0	3
*Sp*Cas9:dsDNA (TGC)	1135	85.7 × 81.4 × 77.3	49507	14939	40	0	3
dsDNA unbound *Sp*Cas9	1135	85.7 × 81.4 × 66.9	42909	13011	17	0	3
*Sp*Cas9:dsDNA (TGG)	1111	75.2 × 81.9 × 72.3	40330	11968	37	0	3
*Sp*Cas9:dsDNA (TGA)	1111	75.2 × 81.9 × 72.3	40325	11966	37	0	3
*Sp*Cas9:dsDNA (TGT)	1111	75.2 × 81.9 × 72.3	40331	11968	37	0	3
*Sp*Cas9:dsDNA (TGC)	1111	75.2 × 81.9 × 72.3	40327	11967	37	0	3
dsDNA unbound *Sp*Cas9	1111	74.1 × 81.9 × 63.3	35090	10705	16	0	3
*Sp*Cas9:dsDNA (TGG)	1337	82.3 × 76.8 × 73.4	42720	12780	21	0	3
*Sp*Cas9:dsDNA (TGA)	1337	82.3 × 76.8 × 71.7	41638	12419	21	0	3
*Sp*Cas9:dsDNA (TGT)	1337	82.3 × 76.8 × 73.4	42502	12707	21	0	3
*Sp*Cas9:dsDNA (TGC)	1337	82.3 × 76.8 × 73.4	42513	12711	21	0	3
dsDNA unbound *Sp*Cas9	1337	82.3 × 76.8 × 73.4	43112	13180	0	2	3
*Sp*Cas9^{R1335V,E1219F}:dsDNA (TGG)	1337	81.9 × 76.8 × 73.4	42117	12580	21	0	3
*Sp*Cas9^{R1335V,E1219F}:dsDNA (TGA)	1337	81.9 × 76.8 × 73.6	42388	12670	21	0	3

*Sp*Cas9^{R1335V,E1219F}:ds DNA (TGT)	1337	81.9 × 76.8 × 73.6	41935	12519	21	0	3
*Sp*Cas9^{R1335V,E1219F}:ds DNA (TGC)	1337	81.9 × 76.8 × 74.8	42717	12780	21	0	3
dsDNA unbound *Sp*Cas9^{R1335V,E1219F}	1337	81.9 × 72.3 × 73.0	40166	12261	0	6	3
*Sp*Cas9:dsDNA (TGG)	1218	76.9 × 76.8 × 74.9	40447	11943	25	0	3
*Sp*Cas9:dsDNA (TGA)	1218	76.9 × 76.8 × 74.9	40070	11817	25	0	3
*Sp*Cas9:dsDNA (TGT)	1218	76.9 × 76.8 × 74.9	40436	11939	25	0	3
*Sp*Cas9:dsDNA (TGC)	1218	76.9 × 76.8 × 74.9	40117	11833	25	0	3
dsDNA unbound *Sp*Cas9	1218	76.9 × 76.8 × 73.2	39905	12033	2	0	3
*Sp*Cas9:dsDNA (TGG)	1322	79.7 × 76.2 × 78.3	43682	12875	34	0	3
*Sp*Cas9:dsDNA (TGA)	1322	79.7 × 76.2 × 78.3	43689	12877	34	0	3
*Sp*Cas9:dsDNA (TGT)	1322	79.7 × 76.2 × 78.3	43692	12878	34	0	3
*Sp*Cas9:dsDNA (TGC)	1322	79.7 × 76.2 × 78.3	43688	12877	34	0	3
dsDNA unbound *Sp*Cas9	1322	79.7 × 72.9 × 65.7	34871	10232	9	0	3
*Sp*Cas9:dsDNA (TGG)	G2 of PAM	86.4 × 98.6 × 88.3	69199	20104	57	0	1
*Sp*Cas9:dsDNA (TGA)	G2 of PAM	86.4 × 98.6 × 88.3	68843	19985	57	0	1
*Sp*Cas9:dsDNA (TGT)	G2 of PAM	86.4 × 98.6 × 88.3	69203	20105	57	0	1
*Sp*Cas9:dsDNA (TGC)	G2 of PAM	86.4 × 98.6 × 88.3	68593	19902	57	0	1
dsDNA unbound *Sp*Cas9	G2 of PAM	86.4 × 98.6 × 88.1	69512	20545	28	0	1

Table A3.2. Alchemical free energy calculations: Residues alchemically transformed, Details on simulation systems, protocol followed, simulation time (in ns) and ΔG estimates (in kcal/mol). Computed standard errors were reported after '±' symbol.

Alchemical Transformation	**System**	**Trial**	**Number of windows × ns per window**	**Run length (ns)**	**ΔG (FEP) (kcal/mol)**	**Average**
R1335 → V1335	***Sp*Cas9:dsDNA (TGG)**	1	51 × 5	250	282.58 ± 0.18	282.63 ± 0.24
		2	51 × 3	150	282.60 ± 0.17	
		3	51 × 3	150	282.14 ± 0.18	
		4 (Full system)	51 × 3	150	283.19 ± 0.22	
		5 (PDB 4UN3)	51 × 3	150	282.80 ± 0.19	282.80 ± 0.19
	***Sp*Cas9:dsDNA (TGA)**	1	51 × 5	250	277.59 ± 0.18	277.91 ± 0.16
		2	51 × 3	150	278.04 ± 0.16	
		3	51 × 3	150	278.11 ± 0.18	
		1	51 × 5	250	270.40 ± 0.14	

	***Sp*Cas9:dsDNA (TGT)**	2	51 × 3	150	270.96 ± 0.25	270.76 ± 0.18
		3	51 × 3	150	270.92 ± 0.16	
	***Sp*Cas9:dsDNA (TGC)**	1	51 × 5	250	271.75 ± 0.15	271.34 ± 0.21
		2	51 × 3	150	271.24 ± 0.16	
		3	51 × 3	150	271.03 ± 0.18	
	dsDNA unbound *Sp*Cas9	1	51 × 5	250	269.87 ± 0.21	270.36 ± 0.26
		2	51 × 3	150	270.22 ± 0.23	
		3	51 × 3	150	270.39 ± 0.16	
		4 (Full system)	51 × 3	150	270.97 ± 0.12	
E1219 → F1219	***Sp*Cas9:dsDNA (TGG)**	1	51 × 5	250	121.88 ± 0.13	121.28 ± 0.37
		2	51 × 3	150	121.36 ± 0.21	
		3	51 × 3	150	120.61 ± 0.14	
	***Sp*Cas9:dsDNA (TGA)**	1	51 × 5	250	119.58 ± 0.12	119.47 ± 0.11
		2	51 × 3	150	119.26 ± 0.16	
		3	51 × 3	150	119.57 ± 0.19	
	***Sp*Cas9:dsDNA (TGT)**	1	51 × 5	250	119.18 ± 0.10	119.48 ± 0.16
		2	51 × 3	150	119.59 ± 0.15	
		3	51 × 3	150	119.68 ± 0.16	
	***Sp*Cas9:dsDNA (TGC)**	1	51 × 5	250	117.58 ± 0.11	118.32 ± 0.38
		2	51 × 3	150	118.80 ± 0.28	
		3	51 × 3	150	118.59 ± 0.23	
	dsDNA unbound *Sp*Cas9	1	51 × 5	250	123.98 ± 0.15	123.49 ± 0.25
		2	51 × 3	150	123.22 ± 0.21	
		3	51 × 3	150	123.26 ± 0.2	
D1135 →V1135	***Sp*Cas9:dsDNA (TGG)**	1	51 × 5	250	129.16 ± 0.13	129.55 ± 0.29
		2	51 × 3	150	129.36 ± 0.18	
		3	51 × 3	150	130.12 ± 0.19	
	***Sp*Cas9:dsDNA (TGA)**	1	51 × 5	250	128.18 ± 0.20	128.74 ± 0.28
		2	51 × 3	150	129.11 ± 0.21	
		3	51 × 3	150	128.92 ± 0.18	
	***Sp*Cas9:dsDNA (TGT)**	1	51 × 5	250	128.01 ± 0.17	128.31 ± 0.16
		2	51 × 3	150	128.43 ± 0.17	
		3	51 × 3	150	128.50 ± 0.18	
	***Sp*Cas9:dsDNA (TGC)**	1	51 × 5	250	130.38 ± 0.17	129.81 ± 0.39
		2	51 × 3	150	129.98 ± 0.16	
		3	51 × 3	150	129.06 ± 0.21	
		1	51 × 5	250	132.09 ± 0.10	

	dsDNA unbound *Sp*Cas9	2	51 × 3	150	132.47 ± 0.13	132.61 ± 0.35
		3	51 × 3	150	133.29 ± 0.12	
L1111 →R1111	***Sp*Cas9:dsDNA (TGG)**	1	51 × 5	250	-256.71 ± 0.15	-257.42 ± 0.36
		2	51 × 3	150	-257.88 ± 0.16	
		3	51 × 3	150	-257.63 ± 0.16	
	***Sp*Cas9:dsDNA (TGA)**	1	51 × 5	250	-256.66 ± 0.19	-256.41 ± 0.17
		2	51 × 3	150	-256.09 ± 0.16	
		3	51 × 3	150	-256.47 ± 0.16	
	***Sp*Cas9:dsDNA (TGT)**	1	51 × 5	250	-256.51 ± 0.17	-256.71 ± 0.22
		2	51 × 3	150	-257.14 ± 0.18	
		3	51 × 3	150	-256.47 ± 0.18	
	***Sp*Cas9:dsDNA (TGC)**	1	51 × 5	250	-257.37 ± 0.18	-257.07 ± 0.17
		2	51 × 3	150	-256.77 ± 0.13	
		3	51 × 3	150	-257.07 ± 0.15	
	dsDNA unbound *Sp*Cas9	1	51 × 5	250	-255.09 ± 0.17	-254.64 ± 0.23
		2	51 × 3	150	-254.49 ± 0.16	
		3	51 × 3	150	-254.35 ± 0.16	
T1337 → R1337	***Sp*Cas9:dsDNA (TGG)**	1	51 × 5	250	-246.32 ± 0.16	-246.30 ± 0.06
		2	51 × 3	150	-246.28 ± 0.18	
		3	51 × 3	150	-246.40 ± 0.25	
	***Sp*Cas9:dsDNA (TGA)**	1	51 × 5	250	-246.98 ± 0.22	-246.50 ± 0.24
		2	51 × 3	150	-246.29 ± 0.16	
		3	51 × 3	150	-246.24 ± 0.14	
	***Sp*Cas9:dsDNA (TGT)**	1	51 × 5	250	-244.37 ± 0.21	-244.43 ± 0.22
		2	51 × 3	150	-244.08 ± 0.18	
		3	51 × 3	150	-244.84 ± 0.15	
	***Sp*Cas9:dsDNA (TGC)**	1	51 × 5	250	-246.87 ± 0.09	-246.39 ± 0.24
		2	51 × 3	150	-246.06 ± 0.12	
		3	51 × 3	150	-246.25 ± 0.11	
	dsDNA unbound *Sp*Cas9	1	51 × 5	250	-244.51 ± 0.16	-244.18 ± 0.17
		2	51 × 3	150	-243.92 ± 0.16	
		3	51 × 3	150	-244.12 ± 0.29	
	***Sp*Cas9^{R1335V,E1219F}: dsDNA (TGG)**	1	51 × 5	250	-249.35 ± 0.14	-249.80 ± 0.36
		2	51 × 3	150	-249.54 ± 0.18	
		3	51 × 3	150	-250.51 ± 0.17	
	***Sp*Cas9^{R1335V,E1219F}: dsDNA (TGA)**	1	51 × 5	250	-247.84 ± 0.22	-247.95 ± 0.22
		2	51 × 3	150	-247.81 ± 0.19	
		3	51 × 3	150	-248.21 ± 0.25	

T1337 → R1337	***Sp*Cas9^{R1335V,E1219F}: dsDNA (TGG)**	1	51 × 5	250	-247.40 ± 0.18	-247.44 ± 0.15
		2	51 × 3	150	-247.31 ± 0.13	
		3	51 × 3	150	-247.61 ± 0.18	
	***Sp*Cas9^{R1335V,E1219F}: dsDNA (TGC)**	1	51 × 5	250	-246.26 ± 0.08	-246.13 ± 0.27
		2	51 × 3	150	-246.51 ± 0.15	
		3	51 × 3	150	-245.62 ± 0.14	
	dsDNA unbound *Sp*Cas9 ^{R1335V,E1219F}	1	51 × 5	250	-244.49 ± 0.17	-244.18 ± 0.16
		2	51 × 3	150	-244.12 ± 0.13	
		3	51 × 3	150	-243.93 ± 0.11	
G12185 →R1218	***Sp*Cas9:dsDNA (TGG)**	1	51 × 5	250	-252.30 ± 0.09	-252.01 ± 0.23
		2	51 × 3	150	-252.15 ± 0.12	
		3	51 × 3	150	-251.56 ± 0.10	
	***Sp*Cas9:dsDNA (TGA)**	1	51 × 5	250	-251.95 ± 0.10	-251.54 ± 0.22
		2	51 × 3	150	-251.21 ± 0.11	
		3	51 × 3	150	-251.46 ± 0.11	
	***Sp*Cas9:dsDNA (TGT)**	1	51 × 5	250	-252.92 ± 0.08	-252.34 ± 0.35
		2	51 × 3	150	-252.39 ± 0.11	
		3	51 × 3	150	-251.70 ± 0.12	
	***Sp*Cas9:dsDNA (TGC)**	1	51 × 5	250	-252.25 ± 0.21	-252.26 ± 0.26
		2	51 × 3	150	-252.71 ± 0.11	
		3	51 × 3	150	-251.82 ± 0.12	
	dsDNA unbound *Sp*Cas9	1	51 × 5	250	-251.98 ± 0.11	-252.02 ± 0.06
		2	51 × 3	150	-252.12 ± 0.15	
		3	51 × 3	150	-251.97 ± 0.11	
A1322→ R1322	***Sp*Cas9:dsDNA (TGG)**	1	51 × 5	250	-266.67 ± 0.14	-267.02 ± 0.23
		2	51 × 3	150	-267.44 ± 0.18	
		3	51 × 3	150	-266.95 ± 0.17	
	***Sp*Cas9:dsDNA (TGA)**	1	51 × 5	250	-266.79 ± 0.14	-267.13 ± 0.22
		2	51 × 3	150	-267.53 ± 0.12	
		3	51 × 3	150	-267.07 ± 0.22	
	***Sp*Cas9:dsDNA (TGT)**	1	51 × 5	250	-266.70 ± 0.13	-266.83 ± 0.15
		2	51 × 3	150	-266.67 ± 0.14	
		3	51 × 3	150	-267.13 ± 0.19	
	***Sp*Cas9:dsDNA (TGC)**	1	51 × 5	250	-266.76 ± 0.14	-267.02 ± 0.39
		2	51 × 3	150	-266.51 ± 0.19	
		3	51 × 3	150	-267.78 ± 0.18	
	dsDNA unbound *Sp*Cas9	1	51 × 5	250	-266.93 ± 0.12	-266.93 ± 0.11
		2	51 × 3	150	-266.74 ± 0.13	

		3	51 × 3	150	-267.11 ± 0.12
All 7 mutations simultaneously	***Sp*Cas9:dsDNA (TGG)**		201 × 3	600	-493.34 ± 0.20
	***Sp*Cas9:dsDNA (TGA)**		201 × 3	600	-500.66 ± 0.24
	***Sp*Cas9:dsDNA (TGT)**		201 × 3	600	-503.15 ± 0.20
	***Sp*Cas9:dsDNA (TGC)**		201 × 3	600	-502.49 ± 0.21
	dsDNA unbound *Sp*Cas9		201 × 3	600	-490.44 ± 0.18

Table A3.3. Trajectory averaged distances (in Å) with standard deviations between key PAM interacting atoms for *Sp*Cas9 and *Sp*Cas9^{R1335V} bound to 5`-TGG-3`, 5`-TGA-3`, 5`-TGT-3` and 5`-TGC-3` PAM sequences from multiple independent trajectories.

R1333, R1335: PAM Interacting partners	**Trials**	***Sp*Cas9: PAM**	***Sp*Cas9 ^{R1335V} PAM**
***Sp*Cas9/*Sp*Cas9^{R1335V} : (TGG) PAM**			
R1333-NH2:G2-N7	1	3.12 ± 0.19	5.58 ± 0.42
	2	2.98 ± 0.12	5.56 ± 0.64
	3	2.96 ± 0.11	5.11 ± 0.28
	Average	**3.02 ± 0.14**	**5.42 ± 0.45**
R1333-NH2:G2-O6	1	3.06 ± 0.33	5.86 ± 0.42
	2	2.79 ± 0.13	5.76 ± 0.60
	3	2.81 ± 0.14	4.58 ± 0.71
	Average	**2.89 ± 0.20**	**5.40 ± 0.58**
***Sp*Cas9/*Sp*Cas9 ^{R1335V} : (TGA) PAM**			
R1333-NH1/2:G2-N7	1	2.95 ± 0.14	4.94 ± 0.69
	2	2.96 ± 0.15	5.02 ± 0.29
	3	2.98 ± 0.13	5.52 ± 0.50
	Average	**2.96 ± 0.14**	**5.16 ± 0.49**
R1333-NH1/2:G2-O6	1	2.85 ± 0.16	5.63 ± 1.36
	2	2.84 ± 0.14	4.89 ± 0.25
	3	2.85 ± 0.17	5.19 ± 0.47
	Average	**2.85 ± 0.16**	**5.24 ± 0.69**
***Sp*Cas9/*Sp*Cas9 ^{R1335V} : (TGT) PAM**			
R1333-NH1/2:G2-N7	1	2.95 ± 0.23	3.04 ± 0.19
	2	2.83 ± 0.55	2.97 ± 0.13
	3	2.97 ± 0.14	2.95 ± 0.23
	Average	**2.92 ± 0.31**	**2.98 ± 0.15**
R1333-NH1/2:G2-O6	1	3.03 ± 0.07	2.84 ± 0.28
	2	2.99 ± 0.15	2.81 ± 0.14
	3	3.58 ± 0.34	3.12 ± 0.32
	Average	**3.20 ± 0.19**	**2.92 ± 0.24**

***Sp*Cas9/*Sp*Cas9 ^{R1335V} : (TGC) PAM**			
R1333-NH2:G2-N7	1	5.30 ± 0.47	4.96 ± 0.61
	2	5.10 ± 0.25	5.50 ± 0.51
	3	5.19 ± 0.36	4.98 ± 1.01
	Average	**5.20 ± 0.36**	5.15 ± 0.71
R1333-NH2:G2-O6	1	4.86 ± 0.21	4.42 ± 0.59
	2	4.90 ± 0.23	4.96 ± 0.27
	3	4.82 ± 0.18	5.08 ± 1.16
	Average	**4.86 ± 0.21**	**4.82 ± 0.66**

Table A3.4. Trajectory averaged distances (in Å) with standard deviations between key PAM interacting atoms for *Sp*Cas9 and *Sp*Cas9^{E1219F} bound to 5`-TGG-3`, 5`-TGA-3`, 5`-TGT-3` and 5`-TGC-3` PAM sequences from multiple independent trajectories.

R1333, R1335: PAM Interacting partners	**Trials**	***Sp*Cas9: PAM**	***Sp*Cas9^{E1219F} PAM**
***Sp*Cas9/*Sp*Cas9^{E1219F} : (TGG) PAM**			
R1333-NH1/2:G2-N7	1	2.99 ± 0.16	2.95 ± 0.13
	2	3.04 ± 0.17	3.02 ± 0.17
	3	2.95 ± 0.11	2.97 ± 0.15
	Average	**2.99 ± 0.14**	**2.98 ± 0.15**
R1333-NH1/2:G2-O6	1	2.80 ± 0.16	2.86 ± 0.18
	2	2.82 ± 0.15	2.87 ± 0.17
	3	2.81 ± 0.12	2.84 ± 0.18
	Average	**2.81 ± 0.14**	**2.86 ± 0.17**
R1335-NH1/2:G3-N7	1	3.03 ± 0.08	2.98 ± 0.15
	2	2.94 ± 0.14	2.82 ± 0.09
	3	2.99 ± 0.14	3.07 ± 0.08
	Average	**2.99 ± 0.12**	**2.95 ± 0.11**
R1335-NH1/2:G3-O6	1	3.06 ± 0.13	2.86 ± 0.17
	2	2.99 ± 0.21	3.05 ± 0.08
	3	2.86 ± 0.14	2.97 ± 0.22
	Average	**2.97 ± 0.16**	**2.96 ± 0.15**
***Sp*Cas9/*Sp*Cas9^{E1219F} : (TGA) PAM**			
R1333-NH1/2:G2-N7	1	2.92 ± 0.11	2.83 ± 0.18
	2	2.94 ± 0.12	3.03 ± 0.18
	3	2.94 ± 0.12	3.02 ± 0.08
	Average	**2.93 ± 0.12**	**2.97 ± 0.14**
R1333-NH1/2:G2-O6	1	2.87 ± 0.18	2.78 ± 0.12

	2	2.86 ± 0.16	2.86 ± 0.20
	3	2.88 ± 0.19	3.07 ± 0.09
	Average	**2.87 ± 0.17**	**2.90 ± 0.13**
R1335-NH1/2:A3-N7	1	5.94 ± 0.66	2.94 ± 0.16
	2	5.75 ± 0.48	3.32 ± 0.37
	3	4.72 ± 0.49	3.08 ± 0.24
	Average	**5.47 ± 0.54**	**3.11 ± 0.26**
R1335-NH1/2:A3-O2P	1	5.93 ± 0.66	3.06 ± 0.08
	2	10.19 ± 0.62	2.82 ± 0.09
	3	6.31 ± 0.96	3.15 ± 0.31
	Average	**7.48 ± 0.75**	**3.01 ± 0.16**
***Sp*Cas9/*Sp*Cas9^{E1219F} : (TGT) PAM**			
R1333-NH1/2:G2-N7	1	2.96 ± 0.12	2.97 ± 0.13
	2	2.91 ± 0.11	2.95 ± 0.12
	3	2.93 ± 0.27	2.91 ± 0.10
	Average	**2.93 ± 0.16**	**2.94 ± 0.12**
R1333-NH1/2:G2-O6	1	2.82 ± 0.14	2.92 ± 0.21
	2	2.84 ± 0.16	2.96 ± 0.22
	3	2.94 ± 0.11	2.92 ± 0.20
	Average	**2.87 ± 0.14**	**2.93 ± 0.21**
R1335-NH1:G2/T3-O2P	1	6.65 ± 1.56	3.38 ± 0.23
	2	5.64 ± 0.44	3.29 ± 0.23
	3	5.43 ± 0.29	3.21 ± 0.21
	Average	**5.91 ± 0.76**	**3.34 ± 0.22**
R1335-NH2:G2/T3-O2P	1	7.75 ± 1.05	3.13 ± 0.30
	2	4.58 ± 0.48	3.13 ± 0.31
	3	7.27 ± 0.30	2.78 ± 0.13
	Average	**6.53 ± 0.61**	**2.89 ± 0.18**
***Sp*Cas9/*Sp*Cas9^{E1219F} : (TGC) PAM**			
R1333-NH1/2:G2-N7	1	4.11 ± 0.93	3.07 ± 0.28
	2	8.05 ± 0.36	3.00 ± 0.21
	3	6.48 ± 0.50	3.08 ± 0.19
	Average	**6.21 ± 0.59**	**3.05 ± 0.23**
R1333-NH1/2:G2-O6	1	5.03 ± 1.29	3.00 ± 0.26
	2	5.12 ± 0.58	3.25 ± 0.40
	3	7.17 ± 0.45	2.87 ± 0.25
	Average	**5.78 ± 0.77**	**3.04 ± 0.30**

R1335-NH2:C3-O2P	1	8.91 ± 2.85	3.38 ± 0.25
	2	13.02 ± 0.71	3.29 ± 0.26
	3	7.88 ± 1.20	3.34 ± 0.15
	Average	**9.94 ± 1.59**	**3.34 ± 0.22**

Table A3.5. Trajectory averaged distances (in Å) with standard deviations between OG atom of S1136 and O3 atom of third PAM nucleotide in *Sp*Cas9 and *Sp*Cas9^{D1135V} from multiple independent trajectories.

Trials	***Sp*Cas9: PAM**	***Sp*Cas9^{D1135V} :PAM**
S1136-OG:G23-O3		
1	4.73 ± 0.52	3.07 ± 0.18
2	3.81 ± 0.47	3.05 ± 0.17
3	4.05 ± 0.40	3.04 ± 0.08
Average	**4.20 ± 0.46**	**3.05 ± 0.14**
S1136-OG:A23-O3		
1	4.47 ± 0.26	3.06 ± 0.26
2	4.04 ± 0.31	3.07 ± 0.17
3	3.98 ± 0.33	2.99 ± 0.23
Average	**4.16 ± 0.30**	**3.04 ± 0.22**
S1136-OG:T23-O3		
1	4.04 ± 0.38	3.05 ± 0.24
2	4.02 ± 0.29	2.98 ± 0.17
3	4.05 ± 0.37	2.86 ± 0.18
Average	**4.04 ± 0.35**	**2.96 ± 0.20**
S1136-OG:C23-O3		
1	4.13 ± 0.33	3.02 ± 0.14
2	4.26 ± 0.30	2.98 ± 0.21
3	4.27 ± 0.34	3.01 ± 0.21
Average	**4.22 ± 0.33**	**3.00 ± 0.19**

Table A3.6. Trajectory averaged distances (in Å) with standard deviations between NH2 atom of R1337 in *Sp*Cas9^{T1337R} and phosphate group (atom: OP2) of third PAM base from multiple independent trajectories.

***Sp*Cas9^{T1337R} : (TGG) PAM**		
R1337-NH2:G3-O2P	1	2.87 ± 0.24
	2	3.02 ± 0.12
	3	2.78 ± 0.11

	Average	**2.89 ± 0.16**
***Sp*Cas9^{T1337R} : (TGA) PAM**		
R1337-NH2:A3-O2P	1	2.82 ± 0.14
	2	3.07 ± 0.19
	3	2.89 ± 0.17
	Average	**2.93 ± 0.17**
***Sp*Cas9^{T1337R} : (TGT) PAM**		
R1337-NH2:T3-O2P	1	6.93 ± 0.53
	2	10.96 ± 0.66
	3	7.04 ± 0.82
	Average	**8.31 ± 0.67**
***Sp*Cas9^{T1337R} : (TGC) PAM**		
R1337-NH2:C3-O2P	1	2.79 ± 0.13
	2	3.08 ± 0.18
	3	2.98 ± 0.20
	Average	2.95 ± 0.17

Table A3.7. Trajectory averaged distances (in Å) with standard deviations between atoms interacting R1337 residue of *Sp*Cas9^{R1335V,E1219F,T1337R} bound to 5`-TGG-3`, 5`-TGA-3`, 5`-TGT-3` and 5`-TGC-3` PAM sequences. from multiple independent trajectories.

***Sp*Cas9^{R1335V,E1219F,T1337R} : (TGG) PAM**		
R1337-NH1/2:G3-O6	1	2.85 ± 0.13
	2	2.75 ± 0.11
	3	2.91 ± 0.32
	Average	**2.84 ± 0.19**
R1337-NH1/2:G3-N7	1	2.95 ± 0.13
	2	3.00 ± 0.17
	3	3.04 ± 0.39
	Average	**2.99 ± 0.23**
***Sp*Cas9^{R1335V,E1219F,T1337R} : (TGA) PAM**		
R1337-NH1/2:A3-N7	1	3.12 ± 0.25
	2	2.95 ± 0.12
	3	3.06 ± 0.20
	Average	**3.04 ± 0.19**

***Sp*Cas9^{R1335V,E1219F,T1337R} : (TGT) PAM**		
R1337-NH1/2:T3-O4	1	3.06 ± 0.36
	2	2.80 ± 0.15
	3	3.23 ± 0.20
	Average	**3.03 ± 0.23**
***Sp*Cas9^{R1335V,E1219F,T1337R} : (TGC) PAM**		
R1337-NH1/2:C3-O2P	1	3.34 ± 0.22
	2	3.04 ± 0.23
	3	3.06 ± 0.11
	Average	**3.15 ± 0.18**

Chapter 4

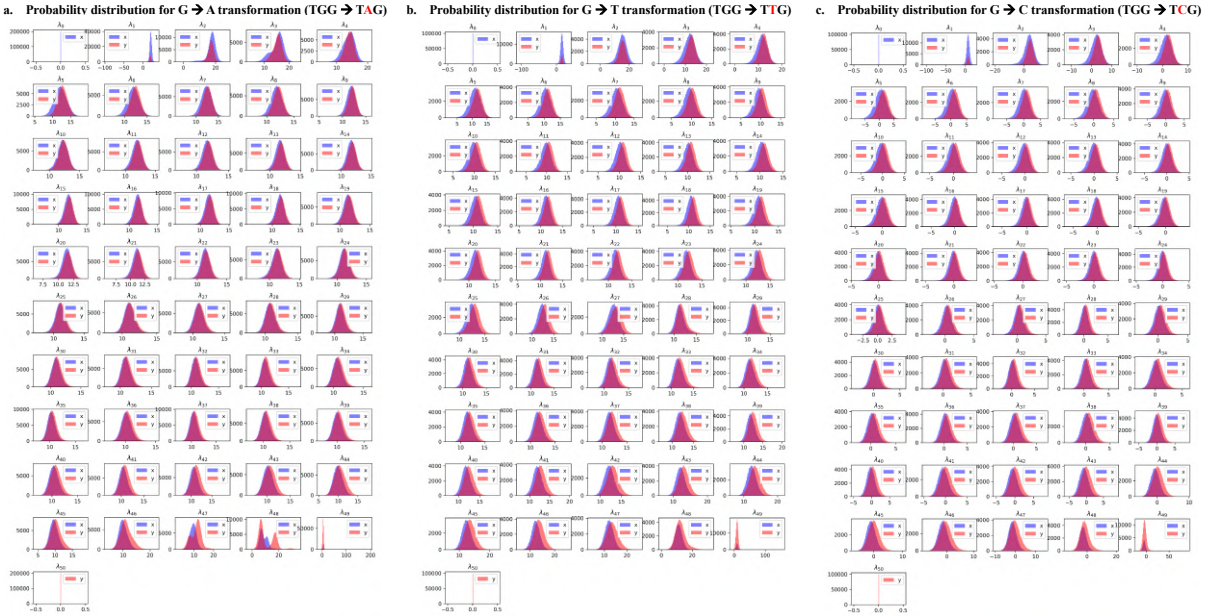


Figure A4.1. Probability distribution functions (Y-axis) depicting forward (blue, x) and reverse (red, y) transformations across all 51 λ windows for: (a) G→A, (b) G→T, (c) G→C. The substantial overlap between forward and reverse distributions ensures the reversibility of the transformations.

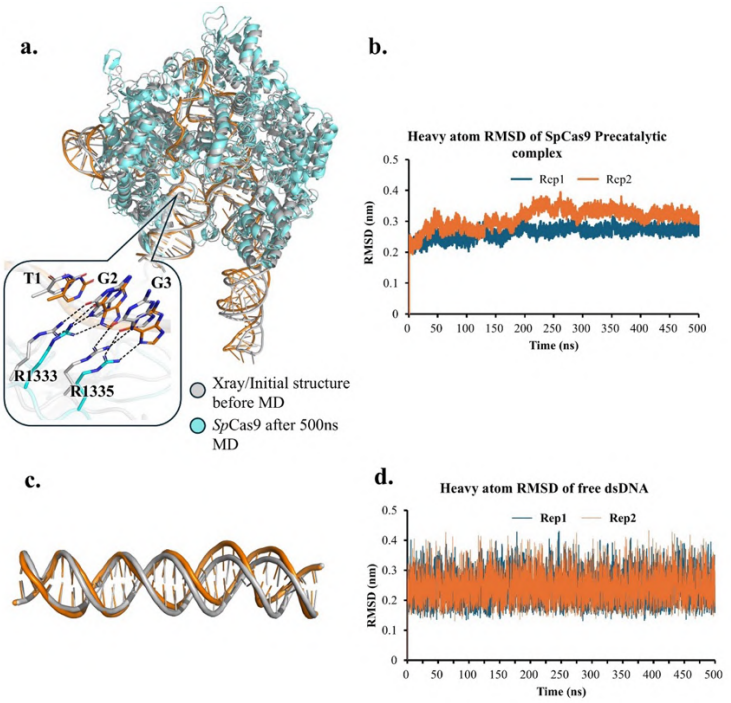


Figure A4.2. (a) Overlay of the X-ray (grey) and MD structure after 500 ns (cyan). The box highlights a zoomed-in view of *Sp*Cas9:PAM. (b) Heavy atom RMSD of the *Sp*Cas9 relative to the X-ray structure for two independent replicates (Rep1: blue, Rep2: orange). (c) Structures of free dsDNA overlaid (initial: grey and final: orange frame of simulation). (d) Heavy atom RMSD of free dsDNA from two independent runs (Rep1 and Rep2) relative to the initial structure.

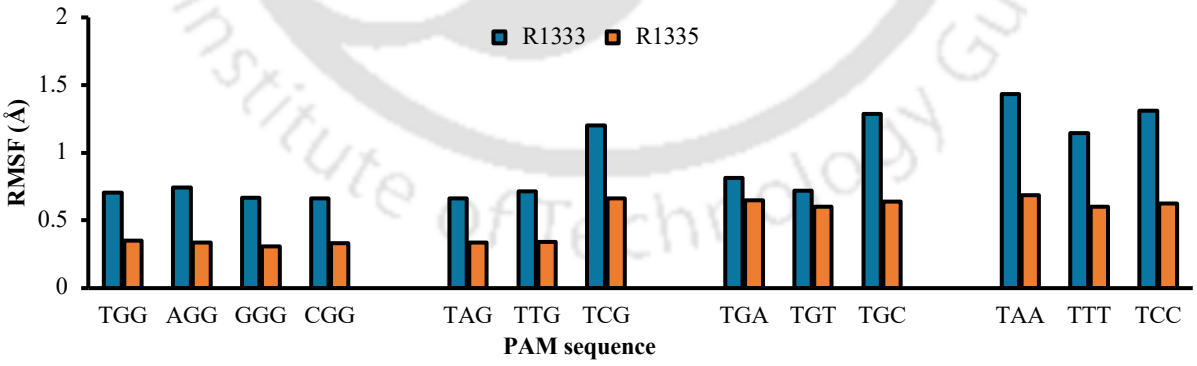


Figure A4.3. Root mean squared fluctuations (RMSF) of two key residues (R1333 and R1335) of *Sp*Cas9 in complex with various dsDNA (canonical: TGG and other non-canonical PAM sequences).

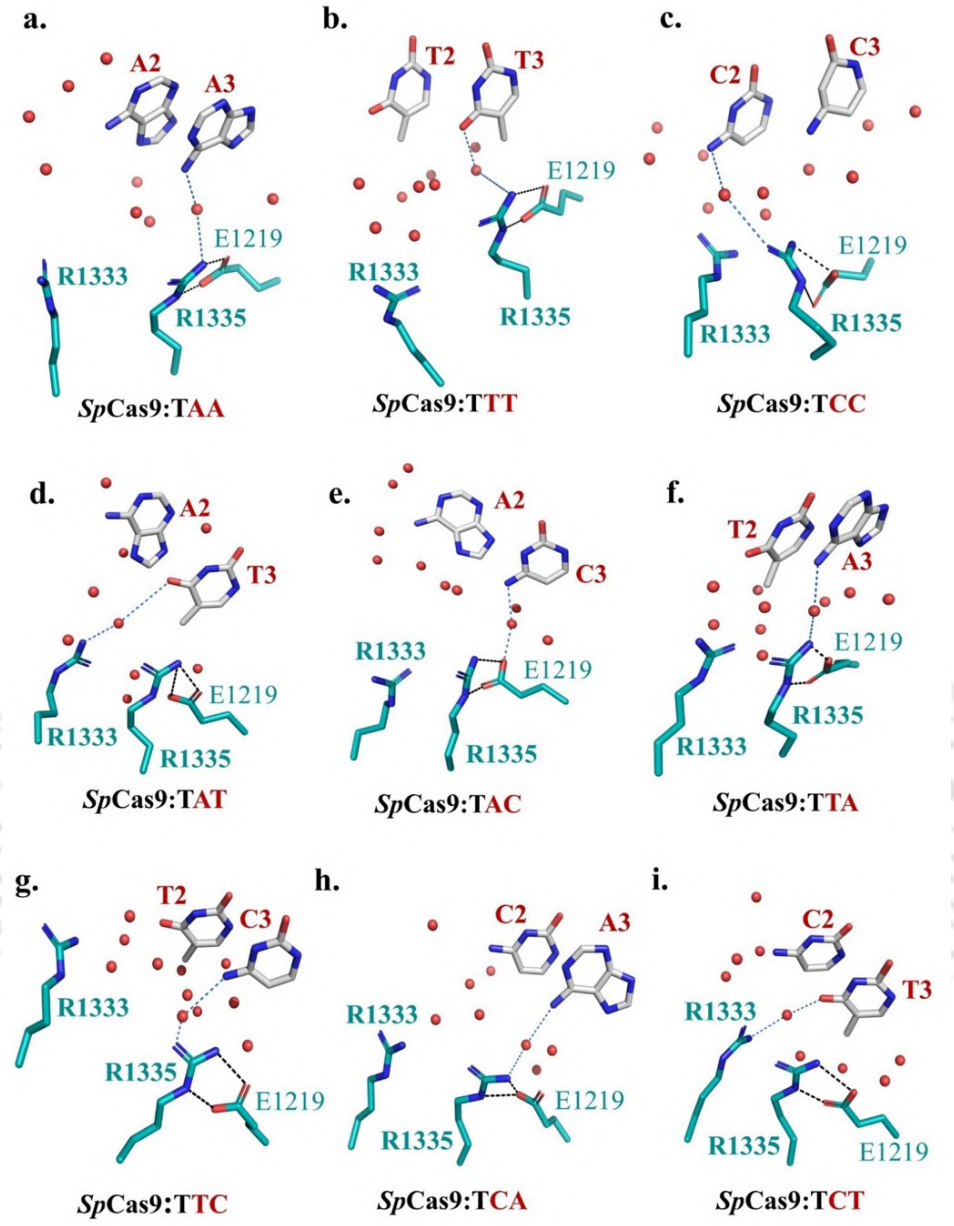


Figure A4.4. Comparison of MD structures of non-canonical double-mutant PAM in complex with *Sp*Cas9: (a) TAA, (b) TTT, (c) TCC, (d) TAT, (e) TAC, (f) TTA, (g) TTC, (h) TCA, (i) TCT. Mutations are shown in red. Key residues are represented in sticks. Direct interactions and water-mediated interactions are depicted as black and blue dotted lines, respectively. Water molecules within 3.5 Å of the Hoogsteen edge of the second and third bases are shown as red spheres. Hydrogen atoms and residue main chains are omitted for clarity.

Table A4.1. Overview of simulation models: system size, box dimension, solvation, and counter ions used.

Simulation model	Box Dimensions (nm^3)	Total no. of atoms	No. of waters	Number of Na^+
For Classical MD simulations				
*Sp*Cas9:sgRNA:dsDNA (TGG)	11.25 × 13.10 × 15.12	217178	63155	135
Free dsDNA (TGG)	12.56 × 12.56 × 12.56	194830	64920	58
For Alchemical free energy simulations (Hybrid topologies)				
*Sp*Cas9:sgRNA:dsDNA (AGG)	11.25 × 13.10 × 15.12	217191	63150	135
Free dsDNA (AGG)	12.56 × 12.56 × 12.56	194840	64284	58
*Sp*Cas9:sgRNA:dsDNA (GGG)	11.25 × 13.10 × 15.12	217202	63154	135
Free dsDNA (GGG)	12.56 × 12.56 × 12.56	194848	64287	58
*Sp*Cas9:sgRNA:dsDNA (CGG)	11.25 × 13.10 × 15.12	217164	63148	135
Free dsDNA (CGG)	12.56 × 12.56 × 12.56	194840	64291	58
*Sp*Cas9:sgRNA:dsDNA (TAG)	11.25 × 13.10 × 15.12	217180	63153	135
Free dsDNA (TAG)	12.56 × 12.56 × 12.56	194841	64291	58
*Sp*Cas9:sgRNA:dsDNA (TTG)	11.25 × 13.10 × 15.12	217197	63152	135
Free dsDNA (TTG)	12.56 × 12.56 × 12.56	194846	64286	58
*Sp*Cas9:sgRNA:dsDNA (TCG)	11.25 × 13.10 × 15.12	217211	63157	135
Free dsDNA (TCG)	12.56 × 12.56 × 12.56	194845	64286	58
*Sp*Cas9:sgRNA:dsDNA (TGA)	11.25 × 13.10 × 15.12	217168	63149	135
Free dsDNA (TGA)	12.56 × 12.56 × 12.56	194838	64290	58
*Sp*Cas9:sgRNA:dsDNA (TGT)	11.25 × 13.10 × 15.12	217197	63152	135
Free dsDNA (TGT)	12.56 × 12.56 × 12.56	194855	64289	58
*Sp*Cas9:sgRNA:dsDNA (TGC)	11.25 × 13.10 × 15.12	217202	63154	135
Free dsDNA (TGC)	12.56 × 12.56 × 12.56	194851	64288	58
*Sp*Cas9:sgRNA:dsDNA (TAA)	11.25 × 13.10 × 15.12	217188	63153	135
Free dsDNA (TAA)	12.56 × 12.56 × 12.56	194834	64286	58
*Sp*Cas9:sgRNA:dsDNA (TTT)	11.25 × 13.10 × 15.12	217237	63156	135
Free dsDNA (TTT)	12.56 × 12.56 × 12.56	194916	64300	58
*Sp*Cas9:sgRNA:dsDNA (TCC)	11.25 × 13.10 × 15.12	217253	63162	135
Free dsDNA (TCC)	12.56 × 12.56 × 12.56	194926	64304	58
*Sp*Cas9:sgRNA:dsDNA (TAT)	11.25 × 13.10 × 15.12	217202	63151	135
Free dsDNA (TAT)	12.56 × 12.56 × 12.56	194872	64292	58
*Sp*Cas9:sgRNA:dsDNA (TAC)	11.25 × 13.10 × 15.12	217204	63152	135
Free dsDNA (TAC)	12.56 × 12.56 × 12.56	194880	64295	58
*Sp*Cas9:sgRNA:dsDNA (TTA)	11.25 × 13.10 × 15.12	217208	63153	135
Free dsDNA (TTA)	12.56 × 12.56 × 12.56	194866	64290	58
*Sp*Cas9:sgRNA:dsDNA (TTC)	11.25 × 13.10 × 15.12	217248	63160	135
Free dsDNA (TTC)	12.56 × 12.56 × 12.56	194921	64302	58
*Sp*Cas9:sgRNA:dsDNA (TCA)	11.25 × 13.10 × 15.12	217210	63154	135
Free dsDNA (TCA)	12.56 × 12.56 × 12.56	194880	64295	58

*Sp*Cas9:sgRNA:dsDNA (TCT)	11.25 × 13.10 × 15.12	217251	63161	135
Free dsDNA (TCT)	12.56 × 12.56 × 12.56	194921	64302	58

Table A4.2. Alchemical free energy calculations: residue transformed, protocols employed, total simulation time (ns), and estimated free energy changes (ΔG_{comp}, ΔG_{free}, $\Delta\Delta G$, kcal/mol). Standard errors are indicated following the '±' symbol.

PAM sequence transformed	Trial	Number of windows × ns per window	Run length (ns) (ΔG_{comp} + ΔG_{free})	ΔG_{comp} (kcal/mol)	ΔG_{free} (kcal/mol)	$\Delta\Delta G$ = ΔG_{comp} - ΔG_{free}(kcal/mol)	Average $\Delta\Delta G$ (kcal/mol)
TGA	1	51 × 3	306	143.41 ± 0.18	134.66 ± 0.11	8.75 ± 0.21	8.63 ± 0.20
	2	51 × 3	306	143.13 ± 0.25	134.83 ± 0.11	8.30 ± 0.27	
	3	51 × 5	510	143.26 ± 0.10	134.42 ± 0.09	8.84 ± 0.13	
TGT	1	51 × 3	306	143.90 ± 0.28	134.91 ± 0.15	9.15 ± 0.32	9.07 ± 0.29
	2	51 × 3	306	143.45 ± 0.27	134.69 ± 0.15	8.76 ± 0.30	
	3	51 × 5	510	144.04 ± 0.26	134.56 ± 0.13	9.48 ± 0.29	
	4	51 × 10	1020	144.27 ± 0.24	134.39 ± 0.11	8.88 ± 0.26	
TGC	1	51 × 3	306	13.07 ± 0.30	1.84 ± 0.16	11.23 ± 0.34	11.34 ± 0.26
	2	51 × 3	306	13.49 ± 0.21	1.67 ± 0.14	11.82 ± 0.25	
	3	51 × 5	510	12.78 ± 0.17	1.79 ± 0.13	10.99 ± 0.21	
TAG	1	51 × 3	306	138.65 ± 0.24	134.26 ± 0.16	4.39 ± 0.29	4.72 ± 0.28
	2	51 × 3	306	139.01 ± 0.26	134.04 ± 0.17	4.6 ± 0.31	
	3	51 × 5	510	139.57 ± 0.22	134.41 ± 0.14	5.16 ± 0.26	
TTG	1	51 × 3	306	141.14 ± 0.21	134.10 ± 0.15	7.04 ± 0.26	7.16 ± 0.23
	2	51 × 3	306	141.58 ± 0.17	134.02 ± 0.14	7.56 ± 0.22	
	3	51 × 5	510	141.51 ± 0.13	134.62 ± 0.17	6.89 ± 0.21	
TCG	1	51 × 3	306	9.76 ± 0.22	.-1.25 ± 0.09	11.01 ± 0.24	11.13 ± 0.24
	2	51 × 3	306	9.99 ± 0.27	.-1.46 ± 0.08	11.45 ± 0.28	
	3	51 × 5	510	9.66 ± 0.17	.-1.28 ± 0.12	10.94 ± 0.21	
AGG	1	51 × 3	306	0.35 ± 0.11	.-0.10 ± 0.12	0.45 ± 0.16	0.25 ± 0.12
	2	51 × 3	306	0.49 ± 0.12	0.06 ± 0.03	0.43 ± 0.12	
	3	51 × 5	510	.-0.46 ± 0.07	.-0.34 ± 0.06	.-0.12 ± 0.09	
CGG	1	51 × 3	306	.-130.54 ± 0.11	.-129.96 ± 0.07	.-0.58 ± 0.13	.-0.46 ± 0.15
	2	51 × 3	306	.-130.23 ± 0.15	.-129.45 ± 0.09	.-0.78 ± 0.17	
	3	51 × 5	510	.-130.82 ± 0.13	.-130.78 ± 0.07	.-0.04 ± 0.15	
GGG	1	51 × 3	306	.-133.86 ± 0.14	.-134.37 ± 0.11	0.51 ± 0.18	0.52 ± 0.19
	2	51 × 3	306	.-134.11 ± 0.15	.-134.57 ± 0.14	0.46 ± 0.21	
	3	51 × 5	510	.-133.64 ± 0.13	.-134.23 ± 0.13	0.59 ± 0.18	
TAA	1	51 × 5	510	290.03 ± 0.19	278.97 ± 0.21	11.06 ± 0.28	11.03 ± 0.41
	2	51 × 5	510	289.74 ± 0.47	279.07 ± 0.29	10.67 ± 0.51	

	3	51 × 5	510	289.83 ± 0.27	278.47 ± 0.35	11.36 ± 0.44	
TTT	1	51 × 5	510	282.06 ± 0.25	270.84 ± 0.27	11.22 ± 0.37	11.43 ± 0.39
	2	51 × 5	510	282.30 ± 0.27	271.27 ± 0.24	11.03 ± 0.36	
	3	51 × 5	510	282.65 ± 0.14	270.59 ± 0.42	12.06 ± 0.44	
TCC	1	51 × 5	510	11.36 ± 0.26	.-0.71 ± 0.25	12.07 ± 0.36	11.64 ± 0.28
	2	51 × 5	510	11.79 ± 0.20	0.24 ± 0.19	11.55 ± 0.27	
	3	51 × 5	510	11.38 ± 0.16	0.08 ± 0.12	11.30 ± 0.20	
TAT	1	51 × 3	306	287.27 ± 0.22	275.53 ± 0.25	11.74 ± 0.33	11.77 ± 0.30
	2	51 × 3	306	287.95 ± 0.31	276.29 ± 0.17	11.66 ± 0.35	
	3	51 × 5	510	288.65 ± 0.19	276.75 ± 0.11	11.90 ± 0.22	
TAC	1	51 × 3	306	148.68 ± 0.30	137.85 ± 0.17	10.83 ± 0.34	11.12 ± 0.34
	2	51 × 3	306	149.15 ± 0.34	137.92 ± 0.22	11.23 ± 0.40	
	3	51 × 5	510	148.96 ± 0.23	137.67 ± 0.14	11.29 ± 0.27	
TTA	1	51 × 3	306	279.93 ± 0.22	268.60 ± 0.21	11.33 ± 0.30	11.20 ± 0.37
	2	51 × 3	306	279.78 ± 0.37	268.85 ± 0.25	10.93 ± 0.44	
	3	51 × 5	510	279.41 ± 0.32	268.06 ± 0.15	11.35 ± 0.36	
TTC	1	51 × 3	306	145.49 ± 0.22	133.88 ± 0.30	11.61 ± 0.37	11.41 ± 0.38
	2	51 × 3	306	146.28 ± 0.41	134.48 ± 0.23	11.8 ± 0.47	
	3	51 × 5	510	145.04 ± 0.19	134.20 ± 0.22	10.84 ± 0.29	
TCA	1	51 × 3	306	145.30 ± 0.29	134.38 ± 0.24	10.92 ± 0.32	11.13 ± 0.36
	2	51 × 3	306	145.32 ± 0.24	134.69 ± 0.35	10.63 ± 0.42	
	3	51 × 5	510	145.84 ± 0.20	133.99 ± 0.26	11.85 ± 0.33	
TCT	1	101 × 3	606	141.60 ± 0.15	130.35 ± 0.12	11.25 ± 0.19	11.14 ± 0.24
	2	51 × 3	306	141.77 ± 0.28	130.94 ± 0.17	10.83 ± 0.32	
	3	51 × 5	510	142.17 ± 0.25	130.65 ± 0.14	11.52 ± 0.29	
	4	51 × 10	1020	141.85 ± 0.05	130.59 ± 0.13	10.96 ± 0.14	

Table A4.3. Salt-bridge occupancy percentages between E1219 and R1335 across different PAM variants and trials. The table summarizes the stability of salt-bridge interactions across multiple molecular dynamics trials for four PAM sequences (TGG, TGA, TGT, TGC). Occupancy values reflect the percentage of simulation frames where the interatomic distance remained below 3.4 Å.

PAM	**Interacting partners**	**Trial**	**Salt bridge Occupancy %**
TGG	E1219-OE1/2: R1335-NH2	trial 1	100.00%
		trial 2	98.06%
		trial 3	90.87%
	E1219-OE1/2: R1335-NE	trial 1	92.65%
		trial 2	97.34%
		trial 3	90.56%
TGA	E1219-OE1/2: R1335-NH2	trial 1	85.34%
		trial 2	91.43%

		trial 3	87.86%
	E1219-OE1/2: R1335-NE	trial 1	84.17%
		trial 2	92.53%
		trial 3	85.69%
TGT	E1219-OE1/2: R1335-NH2	trial 1	100.00%
		trial 2	96.02%
		trial 3	97.50%
		trial 4	97.67%
	E1219-OE1/2: R1335-NE	trial 1	77.72%
		trial 2	86.15%
		trial 3	86.57%
		trial 4	88.22%
TGC	E1219-OE1/2: R1335-NH2	trial 1	90.45%
		trial 2	100.00%
		trial 3	86.13%
	E1219-OE1/2: R1335-NE	trial 1	66.40%
		trial 2	75.17%
		trial 3	71.43%

Table A4.4. Trajectory-averaged key interatomic distances (in Å) for various *Sp*Cas9-PAM complexes. Standard deviations (±) are reported as error. Distances exceeding 3.4 Å are highlighted in red, indicating loss of interactions.

PAM	**Interaction**	**Avg Distance (Å)**
TGG	R1333-NH1/2 : G2-N7	2.99 ± 0.15
	R1333-NH1/2 : G2-O6	2.81 ± 0.14
	R1335-NH1/2 : G3-N7	3.03 ± 0.16
	R1335-NH1/2 : G3-O6	2.90 ± 0.16
AGG	R1333-NH1/2 : G2-N7	3.00 ± 0.14
	R1333-NH1/2 : G2-O6	2.83 ± 0.13
	R1335-NH1/2 : G3-N7	3.04 ± 0.15
	R1335-NH1/2 : G3-O6	2.80 ± 0.10
GGG	R1333-NH1/2 : G2-N7	2.91 ± 0.12
	R1333-NH1/2 : G2-O6	2.95 ± 0.16
	R1335-NH1/2 : G3-N7	3.07 ± 0.16
	R1335-NH1/2 : G3-O6	2.78 ± 0.10
CGG	R1333-NH1/2 : G2-N7	2.97 ± 0.09
	R1333-NH1/2 : G2-O6	2.95 ± 0.10
	R1335-NH1/2 : G3-N7	3.10 ± 0.14
	R1335-NH1/2 : G3-O6	2.79 ± 0.09
TAG	R1333-NH1 : A2-N7	3.10 ± 0.17
	R1335-NH1/2 : G3-N7	2.95 ± 0.14
	R1335-NH1/2 : G3-O6	3.00 ± 0.18
TTG	R1333-NH1 : T2-O4	6.82 ± 0.44
	R1333-NH2 : T2-O4	5.84 ± 0.38
	R1335-NH1/2 : G3-N7	3.00 ± 0.14
	R1335-NH1/2 : G3-O6	2.84 ± 0.14
	R1333-NH1/2 : T1-O1/2P	2.85 ± 0.23
TCG	R1333-NH1 : C2-N4	6.83 ± 0.73

	R1333-NH2 : C2-N4	7.87 ± 0.81
	R1335-NH1/2 : G3-O6	5.28 ± 0.42
TGA	R1333-NH1/2 : G2-N7	2.97 ± 0.15
	R1333-NH1/2 : G2-O6	2.94 ± 0.17
	R1335-NH2 : A3-N7	5.20 ± 0.31
TGT	R1333-NH1/2 : G2-N7	2.95 ± 0.13
	R1333-NH1/2 : G2-O6	2.88 ± 0.13
	R1335-NH1 : T3-O4	4.76 ± 0.34
	R1335-NH2 : T3-O4	5.01 ± 0.31
TGC	R1333-NH1/2 : G2-N7	5.61 ± 0.26
	R1333-NH1/2 : G2-O6	5.59 ± 0.29
	R1335-NH1 : C3-N4	6.89 ± 0.35
	R1335-NH2 : C3-N4	6.21 ± 0.36
TAA	R1333-NH1 : A2-N7	7.58 ± 0.59
	R1333-NH2 : A2-N7	7.75 ± 0.75
	R1335-NH1 : A3-N7	6.27 ± 0.46
	R1335-NH2 : A3-N7	6.42 ± 0.40
TTT	R1333-NH1 : T2-O4	6.38 ± 0.73
	R1333-NH2 : T2-O4	6.63 ± 0.54
	R1335-NH1 : T3-O4	4.87 ± 0.38
	R1335-NH2 : T3-O4	5.02 ± 0.46
TCC	R1333-NH1 : C2-N4	7.33 ± 0.49
	R1333-NH2 : C2-N4	7.44 ± 0.52
	R1335-NH1 : C3-N4	6.37 ± 0.46
	R1335-NH2 : C3-N4	6.28 ± 0.54

TAT	R1333-NH1 : A2-N7	7.95 ± 0.56
	R1333-NH2 : A2-N7	6.98 ± 0.43
	R1335-NH1 : T3-O4	7.71 ± 0.49
	R1335-NH2 : T3-O4	7.92 ± 0.48
TAC	R1333-NH1 : A2-N7	7.10 ± 0.69
	R1333-NH2 : A2-N7	8.29 ± 0.68
	R1335-NH1 : C3-N4	6.70 ± 0.53
	R1335-NH2 : C3-N4	5.44 ± 0.33
TTA	R1333-NH1 : T2-O4	7.04 ± 0.51
	R1333-NH2 : T2-O4	6.30 ± 0.60
	R1335-NH1 : A3-N7	7.03 ± 0.50
	R1335-NH2 : A3-N7	5.68 ± 0.35
TTC	R1333-NH1 : T2-O4	6.12 ± 0.57
	R1333-NH2 : T2-O4	6.55 ± 0.70
	R1335-NH1 : C3-N4	5.33 ± 0.57
	R1335-NH2 : C3-N4	5.44 ± 0.51
TCA	R1333-NH1 : C2-N4	6.31 ± 0.72
	R1333-NH2 : C2-N4	7.51 ± 0.83
	R1335-NH1 : A3-N7	7.86 ± 0.42
	R1335-NH2 : A3-N7	9.53 ± 0.42
TCT	R1333-NH1 : C2-N4	5.50 ± 0.63
	R1333-NH2 : C2-N4	6.47 ± 0.58
	R1335-NH1 : T3-O4	5.62 ± 0.53
	R1335-NH2 : T3-O4	4.97 ± 0.45

Publications from Thesis:

1. Shreya Bhattacharya and Priyadarshi Satpati*. (2022). Insights into the mechanism of CRISPR/Cas9 based genome editing from Molecular Dynamics Simulations. ACS Omega, 8, 2, 1817–1837, https://doi.org/10.1021/acsomega.2c05583.

2. Shreya Bhattacharya and Priyadarshi Satpati*. (2024). Why does the E1219V mutation expand T-rich PAM readability in Cas9 from *Streptococcus pyogenes*? J. Chem. Inf. Model, 64, 8, 3237–3247, https://doi.org/10.1021/acs.jcim.3c01515.

3. Shreya Bhattacharya and Priyadarshi Satpati*. (2025). Energetics of Expanded PAM Readability by Engineered Cas9-NG. J. Chem. Inf. Model, 65, 7, 3628–3639, https://doi.org/10.1021/acs.jcim.5c00011.

4. Shreya Bhattacharya, Keshav Goyal, and Priyadarshi Satpati*. (2025). Thermodynamics of PAM recognition by Cas9 of *Streptococcus pyogenes*. J. Chem. Inf. Model, 65, 24, 13328–13337, https://doi.org/10.1021/acs.jcim.5c01934.

5. Shreya Bhattacharya and Priyadarshi Satpati*. (2025) Hydrophobic fine-tuning in Cas9 enhances non-canonical PAM readability in *Streptococcus pyogenes* (under preparation).

Publications under a collaborative project during thesis work:

- Manorama Ghosal, Tatini Rakshit, Shreya Bhattacharya, Sankar Bhattacharyya, Priyadarshi Satpati* and Dulal Senapati* (2024). E-Protein Protonation Titration-induced Single Particle Chemical Force Spectroscopy for Microscopic Understanding and pI Estimation of Infectious DENV. J. Phys. Chem. B, doi: https://doi.org/10.1021/acs.jpcb.4c00057.

Conferences:

- Participated in Care Conference 2025, Molecules to Medicine: Bioengineering for one health, IIT Madras conducted on 4th – 6th December, 2025.

- Thesis Presentation in Research and Industrial Conclave (RIC), 2025, IIT Guwahati, conducted on 10th October 2024 – 12th October, 2025 (**Best Departmental Innovation Award**).

- Poster Presentation on "Effect of E1219V mutation on the Energetics of PAM recognition by Cas9 from *Streptococcus pyogenes*" in Research and Industrial Conclave (RIC), – Integration'2024, IIT Guwahati, conducted on 9th August 2024 – 11th August, 2024. (**Certificate of Appreciation**)

Workshops:

- Ten days international workshop on Molecular Dynamics Simulation Analysis (Advanced) by Decode Life, conducted on 29th February – 10th March, 2024

- 8 Day Technical Hands-On Certificate Training Md Simulations | Insights Of Computational Biophysics Using Gromacs, Namd & Vmd by BDG Lifesciences (21-28th February, 2022)