UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational Approaches to Understand the Design of Adenine Base Editors

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Chemistry

by

Kartik Lakshmi Rallapalli

Committee in charge:

Professor Francesco Paesani, Chair
Professor Suckjoon Jun
Professor Alexis Komor
Professor Akif Tezcan
Professor Wei Wang
Professor John Weare

2022

The dissertation of Kartik Lakshmi Rallapalli is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my loving family, nurturing mentors, and trusted friends, far away and nearby. I would not be typing these words today without your support and encouragement.

EPIGRAPH

The future is dark, which is the best thing the future can be, I think.

—Virginia Woolf

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

It is difficult, maybe even nearly impossible, to adequately value the support all the individuals without whom this body of work could not have materialized.

To Francesco, *O Captain, My Captain!*, I can not thank you enough for taking a chance on me and allowing me to pursue my eccentric ideas under your guidance. Thank you for your incessant support and faith in me, especially when most people did not and when I had lost faith in my own capabilities. While every Ph.D. trajectory is unique and has its own trials and tribulations, I would not have been able to overcome my deficiencies and inhibitions without your mentorship. Thank you for providing me with the interesting interface of computational chemistry and biochemistry to explore during my research training. Thank you for carving an independent researcher out of me. Thank you for allowing me to pursue my interests and passions outside of the confines of a traditional doctoral training. Thank you for making me feel like a valued member of your research group. A friend of mine once told me that, "Ph.D. is not just about learning science. A lot of people do very different types of science, but in the end we all get awarded a Ph.D. That is because it is about learning the philosophy of science and life". Thank you for mentoring me over these years and teaching me your philosophy of science and life. I will try my best to do justice to all your mentorship. I hope to be as rigorous, precise, creative, bold, and committed as you are, in my next scientific steps.

To Alexis, thank you so much for laying the groundwork for my graduate research and in fact influencing even my next career steps, by creating the base editors. They say most scientific discoveries are serendipitous and I believe I can attest to this adage by attributing all my scientific discoveries to you. If it weren't for you joining the Chemistry department as a Professor at the same time as me joining in as a first year graduate student, my research would look very different today. Thank you for entrusting me with responsibility to understand the biomolecules that you helped discover. Thank you for always being a very supportive mentor to me and teaching me all the biochemistry that I now know, through the fascinating world of genome editing.

To my labmates, in both Paesani lab and Komor lab, thank you for putting up with all my naive questions for five years. Special thanks to Brodie Ranzau, for being my experimental consort in crime. Without your scientific temperament and incredible wet-lab skills I would not have been able to develop my own computational skills. To Sifeng Gu, Mallory Evanoff, Dr. Lambros, Dr. Hunter, and Dr. Egan, thank you for being my scientific sounding board through these years. Whenever I get stuck in any scientific problem, I ask myself what would you do if you were in my place, because I perceive you as the ideal scientist that I am constantly striving to become like.

To Prof. Tezcan, thank you for being incredibly supportive and providing me with critical feedback about my research designs. Not all graduate students have the good fortune to interact this freely with their faculty mentors and sharing the sixth floor of Urey hall has definitely been a privilege. In the same vein, Prof. Jun, thank you for your quick check ups on my research and professional developments, during our brief encounters in the hallways. To Prof. Wang and Prof. Weare, I am extremely grateful to have had your scientific insight throughout my time in graduate school as well as for the opportunity to interact with you through the courses that you taught me.

To my friends, thank you for your generosity, kindness, and affection.

To my family, I exist because of you and for you.

Chapter 1 has been adapted from: Rallapalli, K.L. & Komor, A. C. The Design and Application of DNA Editing Enzymes as Base Editors. *Annu. Rev. Biochem. (under review)*

Chapter 2 has reproduced, in part, with permission, from: Rallapalli, K.L., Komor, A.C., Paesani, F. “Computer simulations explain mutation-induced effects on the DNA editing by adenine base editors”, *Sci. Adv.* 6, eaaz2309 (2020). The dissertation author was the primary author on all reprinted materials.

Chapter 3 has reproduced, in part, with permission, from: Rallapalli, K.L., Ranzau, B.L., Ganapathy, K.R., Paesani, F., Komor, A.C. 2022. Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors. *CRISPR J*. 5:294–310. The dissertation author was the primary author on all reprinted materials.

Chapter 5 has been adapted from: Rallapalli, K.L. & Komor, A. C. The Design and Application of DNA Editing Enzymes as Base Editors. *Annu. Rev. Biochem. (under review)*

VITA

2015	B. Sc. in Chemistry, Miranda House, University of Delhi, Delhi
2017	M. Sc. in Chemistry, Indian Institute of Technology, Delhi, Delhi
2022	Ph. D. in Chemistry, University of California San Diego

PUBLICATIONS

Rallapalli, K. L. & Komor, A. C. The Design and Application of DNA Editing Enzymes as Base Editors. *Annu. Rev. Biochem. (under review)*

Rallapalli, K. L., Ranzau, B. L., Ganapathy, K. R., Paesani, F., & Komor, A. C. Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors. *The CRISPR J.* 2022

Rallapalli, K. L., Komor, A. C., & Paesani, F. Computer simulations explain mutation?induced effects on the DNA editing by adenine base editors. *Sci. Adv.* 2020

Fox, K., **Rallapalli, K. L.**, & Komor, A. C. Rewriting human history and empowering indigenous communities with genome editing tools. *Genes* 2020

FIELD OF STUDY

Major Field: Chemistry

Theoretical and Computational Chemistry, Biochemistry Professor Francesco Paesani

ABSTRACT OF THE DISSERTATION

Computational Approaches to Understand the Design of Adenine Base Editors

by

Kartik Lakshmi Rallapalli

Doctor of Philosophy in Chemistry

University of California San Diego, 2022

Professor Francesco Paesani, Chair

The ability to induce desired changes in the genetic code of an organism, in a precise and controlled fashion, is a long-standing ambition of the life sciences. This goal has now been realized through the development of CRISPR-based genome editing enzymes, that have enabled researchers to target and edit the genome with unprecedented precision.

Among the currently available compendium of CRISPR-based genome editors, base editors, are the most promising candidate for curing more than 60% of all known genetic diseases due to there ability to repair individual DNA bases. The base editor technology relies on a Cas9 protein fused to a single-stranded DNA (ssDNA) modifying enzyme to directly covalently modify target nucleobases in genomic DNA.

The most recent base editors, the adenine base editors (ABEs), catalyzes the conversion of A•T→G•C base pairs at precise genomic loci and were developed using extensive protein engineering and evolution starting from a RNA-editing enzyme, TadA. Given its unique trajectory from an RNA-editing enzyme into an efficient and precise DNA-editing base editor and considering that expansion of the current base editing arsenal would require similar engineering efforts, understanding the molecular design principles for base editor design can help accelerate the field of genome editing. This thesis aims to learn these principles through computational simulations as well as bioinformatics analyses of ABEs as the prototypical base editor.

First, we explore the onset of DNA-editing activity in TadA due to the first ever mutation discovered in its evolutionary journey into becoming the ABEs. Using molecular dynamics simulations we uncover the structural and functional roles played by this initial mutation. We demonstrate that this critical first mutation enhances the binding affinity of TadA towards DNA and verify its significance through *in silico* and *in vivo* reversion analyses.

Subsequently, we show that this critical first mutation is capable of enhancing not just the DNA-editing activity of TadA but also its undesirable native RNA-editing functionality. In an attempt to discover mutations that can suppress the native RNA-editing of TadA and to rationalize the effect of all the reported ABE mutations simultaneously, we developed a sequence-based bioinformatics classifier. This classifier relies on evolutionary information learned from naturally-occurring proteins and can aid the laboratory-evolution of novel base editors.

Finally, we determine the significance of the remote mutations that happen far away from the active site of TadA. Using molecular modeling and large-scale simulations of entire ABE-DNA complex we show that these remote mutations modulate the conformation of terminal end of the TadA and thus help broadening its DNA-editing substrate-specificity while also suppressing its native RNA-editing activity.

Chapter 1

1.1 Introduction

The answer to the question ‘What is Life?’ is highly controversial and depends upon as well as varies with the disciplinary background, explanatory context, and temperamental values of the being answering it. For the present purposes it is suffice to say that life is like an alphabet soup. Blurring out the cellular-level details and ignoring the philosophical complexities entirely, leaves us with a molecular definition of life which requires only 20 or so letters of the English alphabet. Although trivialized, this naïve definition allows us to describe the central dogma of molecular biology as merely the transcription of the four letters encoding DNA into the four letters encoding RNA and its subsequent translation into the twenty letters encoding proteins. Collectively, these three macromolecules with their seemingly random combination of the alphabet, encode, sustain, and propagate life as we know it. Additionally, misspellings in this genetic master code can lead to mistakes at the protein-level that impact the life of the organism.

In fact, almost 60% of genetic diseases in humans are attributed to single nucleotide variations(SNVs), that is single letter misspellings or point mutations, in the DNA **(Figure 1.1)**. Correcting and reversing such SNVs is a long-standing ambition of life sciences as it gives us the power to manipulate and rewrite the code of life. This goal has been recently realized through the development of CRISPR-based genome editors, which have enabled researchers to target and edit the genome with unprecedented precision.[1–3] Among the currently available compendium of CRISPR-based genome editors, the base editor enzymes are the most promising candidate for curing diseases that are caused due to SNVs in the genome.[4]

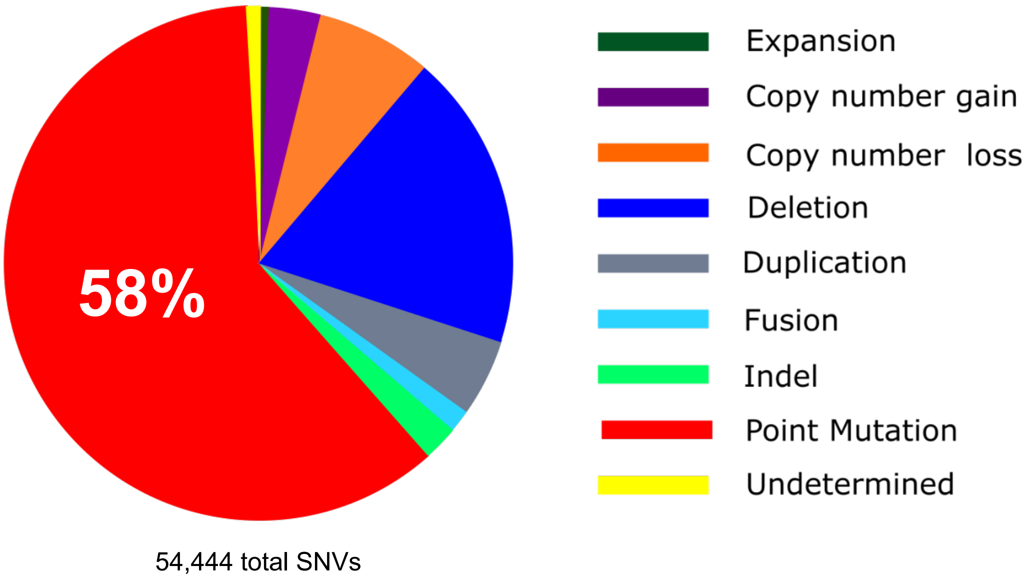


Figure 1.1: **Disease-causing human genetic variations.** a) Distribution of human pathogenic genetic variants in the ClinVar database,[5] emphasizing that the majority of the genetic diseases originate due to SNVs. Adapted from Ref.[4]

.

Base editors rely on a simple modular architecture: a DNA-targeting module, typically a catalytically impaired Cas9, which can recognize specific locations in the genome in a programmable-fashion, fused to a DNA-editing module, which enables the transformation of one nucleobase into another, along with optional accessory modules which enhance the purity of the desired edit (**Figure 1.2**). Overall, base editors leverage the precision and programmability of the DNA-targeting module and combine it with efficient chemical modification capabilities of the DNA-editing module to rewrite the genetic information one-letter at a time[4,6] and have already revolutionized biology - from bench all the way to the clinic.[7–9]

Three classes of genomic DNA base editors have been developed thus far: cytosine base editors(CBEs) that introduce CG→TA point mutations,[10] adenine base editors(ABEs) that

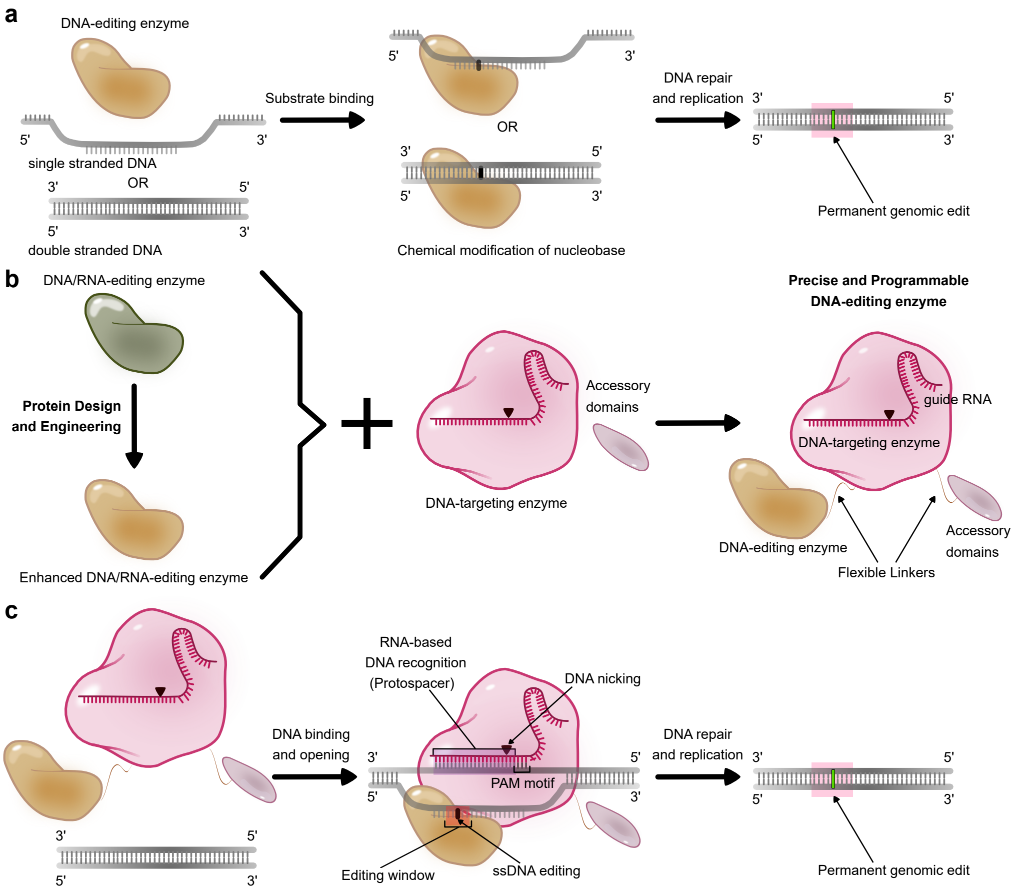


Figure 1.2: **A schematic representation of the design and application of base editors.** (a) DNA-editing enzymes refer to a diverse set of enzymes that bind to DNA and chemically modify a target nucleobase. The resulting modified nucleobase is processed or interpreted by the cellular repair, replication, or transcriptional machinery as an alternate base, resulting in coding or epigenetic changes in the genome. (b) Naturally-occurring DNA editing enzymes as well as enhanced or engineered variants can be combined with precise and programmable DNA-targeting enzymes such as CRISPR-Cas proteins, along with optional accessory domains (which sometimes are DNA editing enzymes themselves) to produce modular enzyme complexes called base editors. (c) The DNA-targeting module of a base editor (usually CRISPR-Cas9) recognizes and binds to the target genomic locus via base-pairing between a guide RNA (gRNA) molecule and the genomic DNA (this region of the DNA is called the protospacer). The protospacer must also be directly next to a protospacer adjacent motif (PAM) to facilitate Cas enzyme binding (for the most commonly used Cas9 system, this PAM sequences is NGG). The Cas protein unwinds the DNA double helix and exposes a small region of single-stranded DNA. If the DNA editing enzyme's substrate is ssDNA, its DNA editing activity is focused on this ssDNA 'editing window', and will chemically modify target nucleobases within this window. If, however, the DNA editing enzyme targets double-stranded DNA, it can modify target nucleobases within the general vicinity of the protospacer. Once processed by the cell's DNA repair and replication machinery, these DNA edits become permanently incorporated into the genome.

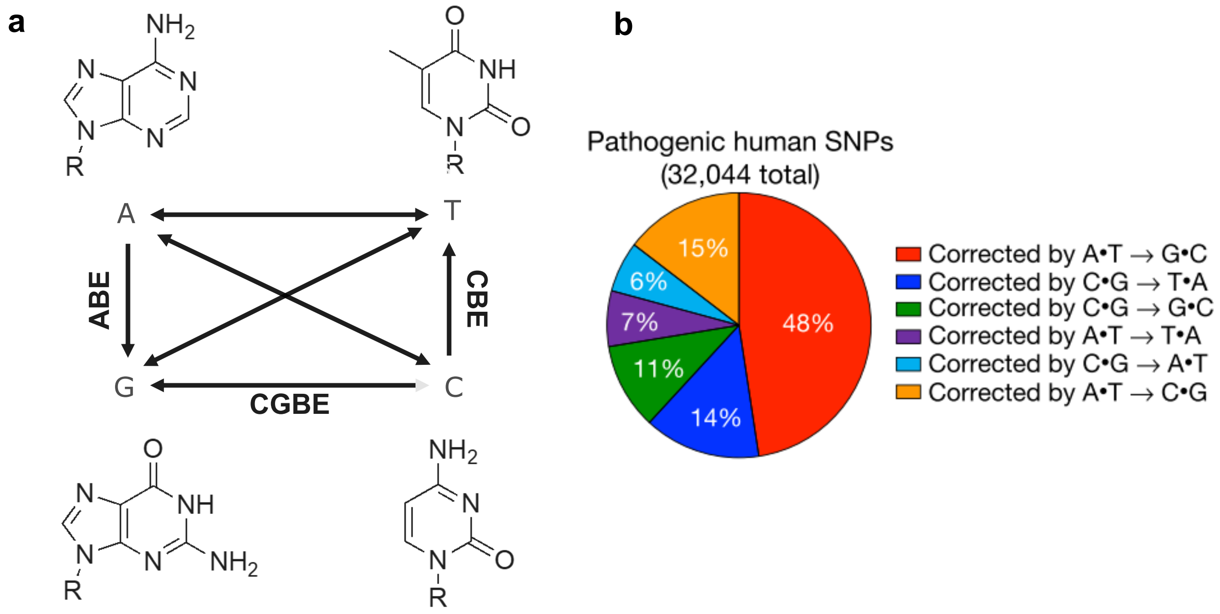


Figure 1.3: **Currently existing base editors and their applicability towards reversing known human pathogenic SNVs.** (a) The four bases of the DNA and all possible interconversions between them. The conversions which are currently achievable using the existing bases editors are highlighted. (b) Distribution of base pair edits needed to reverse the known human pathogenic point mutations. Nearly 50% of SNVs can be cured by A·T→G·C base editing alone. Adapted from Ref.[11]

introduce AT→GC point mutations,[11] and CGBEs that introduce CG→GC point mutations[12–15] (**Figure 1.3(a)**). Of these existing base editors, ABEs, due to their potential to fix nearly 50% of all known the pathogenic SNVs, have garnered significant interest from both academic and clinic researchers. Furthermore, ABEs, unlike CBEs and CGBEs, had to be developed and designing using extensive protein engineering efforts due to the lack of proteins which can chemically catalyze A→G base change in the DNA.

Given the unique and challenging trajectory taken by the ABEs to become an efficient DNA base editors and the incredible power it has as a precise therapeutic technology, in this dissertation we focus here on understanding the design and development of this base editor, specifically with the goal of gaining predictive insights through computational approaches.

1.2 Significance and role of Cas effectors in base editors

Adenine base editor (ABE) adhere to the basic base editor architecture and consists of two modules: a catalytically impaired Cas9, which serves as the DNA-targeting module, and an engineered variant of a tRNA adenosine deaminase (TadA*), which serves as single-stranded DNA(ssDNA)-editing module and enables the transformation of an A:T base pair into a G:C base pair.[11] While this dissertation focuses predominantly on the DNA-editing module of the ABEs, the importance and significance of the DNA-targeting module, that is the Cas9 enzyme cannot be over-emphasized. Hence, before delving deep into the DNA-editing module and its role in the ABEs, we take a brief look at the the Cas enzymes for their role as the DNA-targeting module in the base editors, as after all, it is the discovery and application of Cas-systems which ultimately, albeit unwittingly, led to the development of all base editors. Hence, before delving deep into the DNA-editing module and its role in the ABEs, we take a brief look at the the Cas enzymes. We do this specifically in the light a their role as the DNA-targeting module in the base editors, as after all, it is the discovery and application of Cas-systems which ultimately, albeit unwittingly, led to the development of all base editors.

CRISPR-Cas systems are RNA-guided endonuclease complexes that impart innate immunity to bacteria and archaea against invading phages. Cas effector complexes recognize target nucleic acid sequences via simple base pairing rules between the Cas complex's guide RNA (gRNA) and the target DNA or RNA (**Figure 1.2(c)**). Due to the ease and simplicity with which these systems can be re-targeted to custom DNA and RNA sequences of interest, Cas enzymes have been quickly and universally adopted in the genome editing field as programmable DNA- and RNA-targeting endonucleases. In this review we will focus on DNA-targeting CRISPR-Cas systems. The Cas protein first binds to a gRNA molecule, which is comprised of a "gRNA backbone" (the portion of the gRNA that the Cas protein recognizes and binds to) and a spacer sequence (typically 20-25 bases), which is complementary to the target genomic sequence. The resulting ribonucleoprotein (RNP) complex then searches the genome for protospacer adjacent motifs (PAMs), which are short sequences (typically 2-7 bases) that the Cas protein recognizes and requires for DNA binding. Once the RNP complex finds a PAM that is directly adjacent to a match to the gRNA spacer sequence (the "protospacer"), it will unwind the DNA and base-pair the protospacer with the gRNA spacer (**Figure 1.2(c)**). Binding of the Cas:gRNA complex to the protospacer creates an R-loop, where the "non-target" DNA strand (the strand that is not base-paired with the gRNA) lacks a base-pairing partner. While most of this strand is enveloped within the Cas protein, a subset of this strand is exposed to the surrounding cellular environment. Following successful RNA-mediated binding of the RNP complex to its target DNA sequence, the endonuclease domains of the Cas enzyme will introduce a double-stranded DNA break (DSB) into the target locus. The targeted introduction of a DSB is the first step in "traditional" genome editing, which is followed by cellular processing of the DSB by one of two DNA repair pathways. Processing by non-homologous end-joining (NHEJ) results in the insertion and deletion of nucleotides (collectively referred to as "indels") at the site of the DSB, in an uncontrollable manner. Processing by homology-directed repair (HDR) uses an exogenously supplied "donor DNA" as a template for repair. The donor DNA is designed to have homology to the sequence surrounding the DSB, but encodes an edit of interest. This edit is then incorporated into the genomic DNA during HDR. Typically, both DNA repair pathways compete with one another to process the DSB, resulting in mixtures of genome editing outcomes. "Nontraditional" genome editing methods utilize non-DSB DNA damage product as intermediates, and thus partially or fully catalytically inactivated Cas enzymes are used. In the context of base editing, Cas effectors have the following crucial functions:

- Carrying the DNA editing enzyme to the target genomic locus and localizing the enzymatic activity of the DNA editing enzyme to a nucleobase of interest
- Exposing a "bubble" of ssDNA substrate to the DNA editing enzyme via R-loop formation

(this is important when the DNA editing enzyme's substrate is ssDNA)

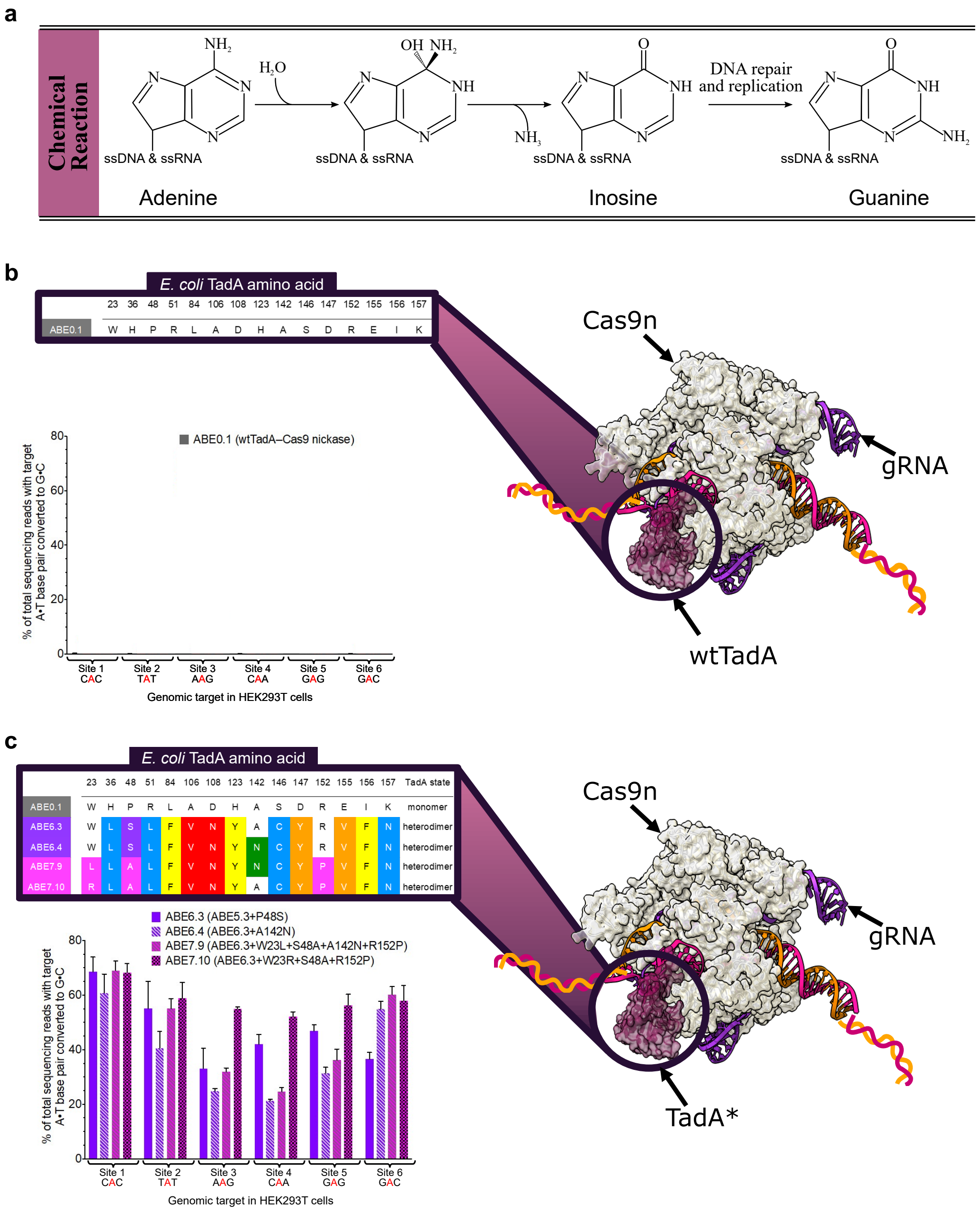
- Nicking the backbone of the DNA strand opposite from the edited target nucleobase, to stimulate DNA repair

1.3 Adenine base editors and their mechanism of action

Once the Cas enzyme has located its target genomic loci and exposed the small 'window' of ssDNA to the nuclear environment, it is the role of the DNA-editing module of the base editor to actually carry out the covalent modification of the DNA base. In the case of the ABEs, this entails the modification of A·T→G·C on the exposed target ssDNA.

While there exist naturally-occurring enzymes which catalyze the transformation of A→G via an inosine(I) intermediate(Figure 1.4(a)), these enzymes can only act on either single-stranded RNA(Tad/ADAT enzymes) or double-stranded RNA(ADAR enzymes),[16] and not on single-stranded DNA(ssDNA). Hence, the prototypical ABE, henceforth referred to as the ABE0.1, which relied on a transfer RNA (tRNA) adenosine deaminase enzyme, TadA as its DNA-editing module, displays no catalytic activity on ssDNA(Figure 1.4(b)), even though RNA and DNA are built out of the similar building blocks(**A**, G, C, (T/U) nucleobases in DNA and RNA).

To circumvent this issue, Gaudelli *et al.* evolved TadA(into TadA*) so that it could accept the ssDNA exposed during the R-loop formation by the Cas9 module as its substrate(Figure 1.2(c)) and successfully conduct A·T→G·C base editing on DNA. Starting from TadA, seven rounds of directed evolution identified 14 amino acid mutations that transformed this RNA editing enzyme into a highly efficient ssDNA editing adenine base editor(ABE7.10)(Figure 1.4(b)). These amino acid mutations, that is changes in the letters that make up the TadA enzyme, that transform it into an efficient DNA-editing enzyme which conduct A→G letter changes at precise locations in the genome is the motivation behind this dissertation(Figure 1.5). Understanding the significance



	23	36	48	51	84	106	108	123	142	146	147	152	155	156	157
ABE0.1	W	H	P	R	L	A	D	H	A	S	D	R	E	I	K

	23	36	48	51	84	106	108	123	142	146	147	152	155	156	157	TadA state
ABE0.1	W	H	P	R	L	A	D	H	A	S	D	R	E	I	K	monomer
ABE6.3	W	L	S	L	F	V	N	Y	A	C	Y	R	V	F	N	heterodimer
ABE6.4	W	L	S	L	F	V	N	Y	N	C	Y	R	V	F	N	heterodimer
ABE7.9	L	L	A	L	F	V	N	Y	N	C	Y	P	V	F	N	heterodimer
ABE7.10	R	L	A	L	F	V	N	Y	A	C	Y	P	V	F	N	heterodimer

Figure 1.4: **Mechanism and development of ABEs.** (a) (b) Basic architecture of ABE0.1 indicating that wtTadA has no measurable A·T→G·C base editing activity for human genomic DNA. (c) Basic architecture of ABE7.10 highlighting the 14 amino acid mutations its DNA-editing module that convert wtTadA into TadA*7.10 and allow it to conduct efficient (averaging 58% across the six genomic sites) A·T→G·C base editing for human genomic DNA. Graphs in (b) and (c) are adapted from Ref.[11]

of these 14 amino acid mutations in TadA is the crucial first step in revealing the design rules necessary to convert any RNA-editing enzyme into a base editor and help in expanding the current genome editing toolkit to include all the possible single letter swaps in the DNA(Figure 1.3(a)).

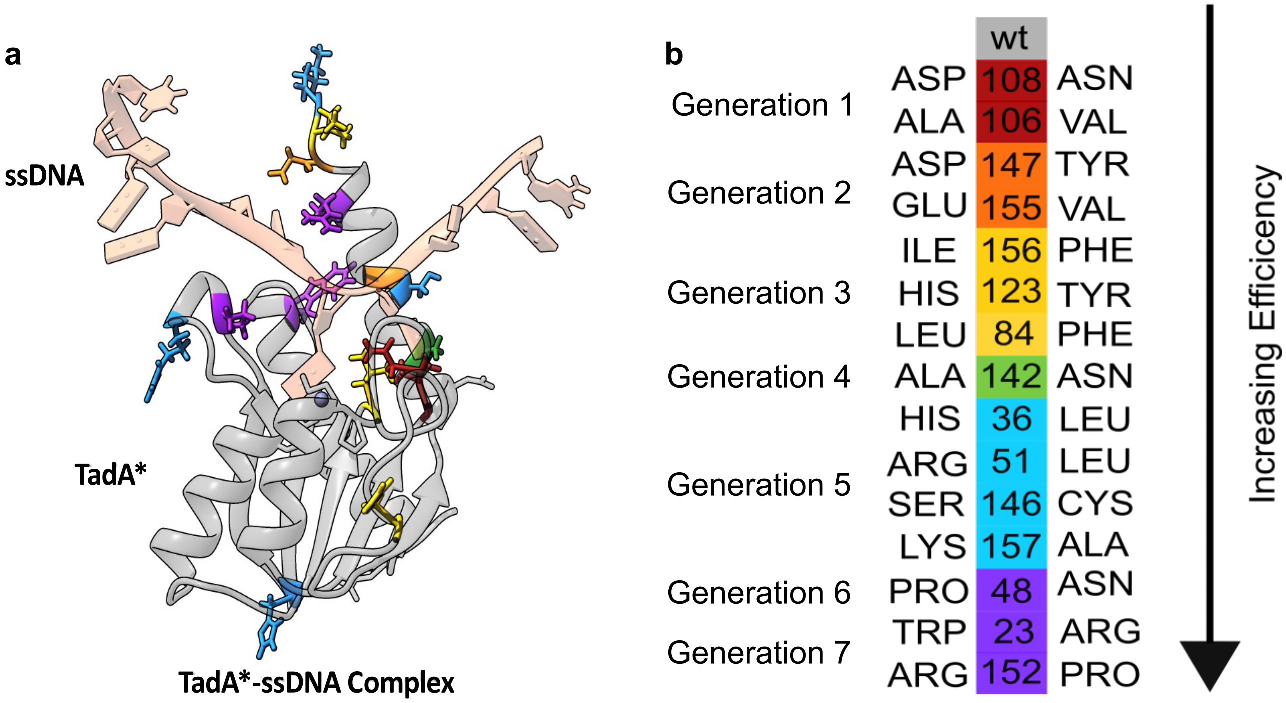


Figure 1.5: **Amino acid mutations that convert RNA-editing wtTadA into DNA-editing TadA7.10.** (a) Model of ecTadA(PDB ID: 1Z3A[17]) bound to substrate ssDNA. The mutations that lead to an increase in its the DNA-editing activity are shown as colored sticks, where the color signifies the evolutionary generation that the mutation was identified during, based on the scheme shown in (b).

1.4 Dissertation objectives

The ultimate aim of this dissertation is to develop a predictive understanding of the ABEs using computational approaches. This is achieved through a following objectives:

- Understanding the onset of DNA-editing activity in ABEs
- Understanding the off-target RNA-editing activity of ABEs
- Understanding the effects of remote mutations in ABEs

Chapter 2

Understanding the onset of DNA-editing activity in ABEs

2.1 Abstract

Adenine base editors, which were developed by engineering a transfer RNA adenosine deaminase enzyme (TadA) into a DNA editing enzyme (TadA*), enable precise modification of A•T→G•C base pairs. Here, we use molecular dynamics simulations to uncover the structural and functional roles played by the initial mutations in the onset of the DNA editing activity by TadA*. Atomistic insights reveal that early mutations lead to intricate conformational changes in the structure of TadA*. In particular, the first mutation, Asp108Asn, induces an enhancement in the binding affinity of TadA to DNA. *In silico* and *in vivo* reversion analyses verify the importance of this single mutation in imparting functional promiscuity to TadA* and demonstrate that TadA* performs DNA base editing as a monomer rather than a dimer.

2.2 Introduction

Base editing is a new genome-editing technology that enables the conversion of one base pair into another at a genomic locus of interest through the precise chemical modification of a target nucleotide.[4,6,10,11] Base editors consist of two subunits: a catalytically impaired Cas9 subunit [Cas9 nickase (Cas9n)] that acts as a DNA binding module and a single-stranded DNA (ssDNA)–specific editing enzyme subunit. The Cas9n binds to a preprogrammed genomic locus and opens the double-stranded DNA to expose a short stretch of ssDNA.[18,19] Subsequently, the ssDNA editing component carries out a chemical reaction to transform a target nucleobase into a noncanonical base (**Figure 2.1**). Last, DNA replication or repair enzymes process the resulting mismatch into a canonical base pair to catalyze an overall base substitution reaction (1). Two types of base editors have been reported to date: cytosine base editors (CBEs), which rely on naturally occurring APOBEC enzymes[20,21] to induce C•G→T•A mutations via a uracil intermediate,[10] and adenosine base editors (ABEs), which use a modified version of the transfer RNA (tRNA) adenosine deaminase enzyme TadA to induce A•T→G•C mutations via an inosine intermediate (**Figure 2.1**).[11] Both editors catalyze a deamination reaction at the target nucleobase and hence display considerable similarity between both the structure and mechanism of their enzymatic subunits.

Since wild-type TadA (wtTadA) was unable to perform adenosine deamination chemistry on ssDNA, despite its structural similarity to several ssDNA modifying enzymes of the APOBEC family,[22] the development of ABEs required extensive protein engineering and evolution efforts. Starting with the TadA enzyme from *Escherichia coli*,[17] which deaminates the wobble position of tRNA^{Arg}, directed evolution[23] was used to achieve efficient editing on a ssDNA substrate. Seven rounds of directed evolution identified 14 point mutations that transformed TadA into ABE7.10, which displays both high editing efficiency and broad sequence compatibility.[11] Understanding the effects of the mutations identified in TadA during the initial rounds of evolution is critical, particularly considering that expansion of the current base editing arsenal would require similar protein engineering and evolution efforts. Evolving enzymes from zero initial activity is notoriously challenging, as it requires screening an enormous sequence space for a select few mutants that impart new activity upon the enzyme of interest; evolution projects that improve upon weak initial activity see higher success rates in contrast.[24] Therefore, a molecular understanding of how the initial TadA mutations gave rise to nonzero DNA editing activity would be indispensable for aiding future evolution efforts. While the wild-type TadA enzyme does not exhibit any enzymatic activity on ssDNA when fused to Cas9n, the first two rounds of identified mutations (Asp108Asn, Ala106Val, Asp147Tyr, and Glu155Val) are responsible for imparting experimentally detectable levels of DNA editing activity to TadA*-Cas9n (* indicates incorporation of mutations).[11] Atomistic understanding of these mutations that cause the onset of detectable activity is paramount to rationally guide the development of future base editors. In this study, we use a combination of molecular dynamics (MD) simulations complemented with experimental measurements to scrutinize the structural and functional implications of these initial mutations.

2.3 Results

2.3.1 Suppression of structural flexibility

We initiated our investigations into the effects of the TadA mutations by studying their influence on the overall structure of the protein. As the first two generations of ABE complexes are composed of a TadA monomer fused to Cas9n (the wild-type enzyme acts on tRNA as a dimer), we furthermore focused our studies on monomeric TadA mutants. In addition, while the final generation ABE7.10 construct is composed of a wtTadA-TadA* dimer fused to Cas9n, we measured the A•T→G•C base editing efficiency of the monomeric TadA7.10*-Cas9n construct at six different target As in human embryonic kidney (HEK) 293T cells and found no decrease

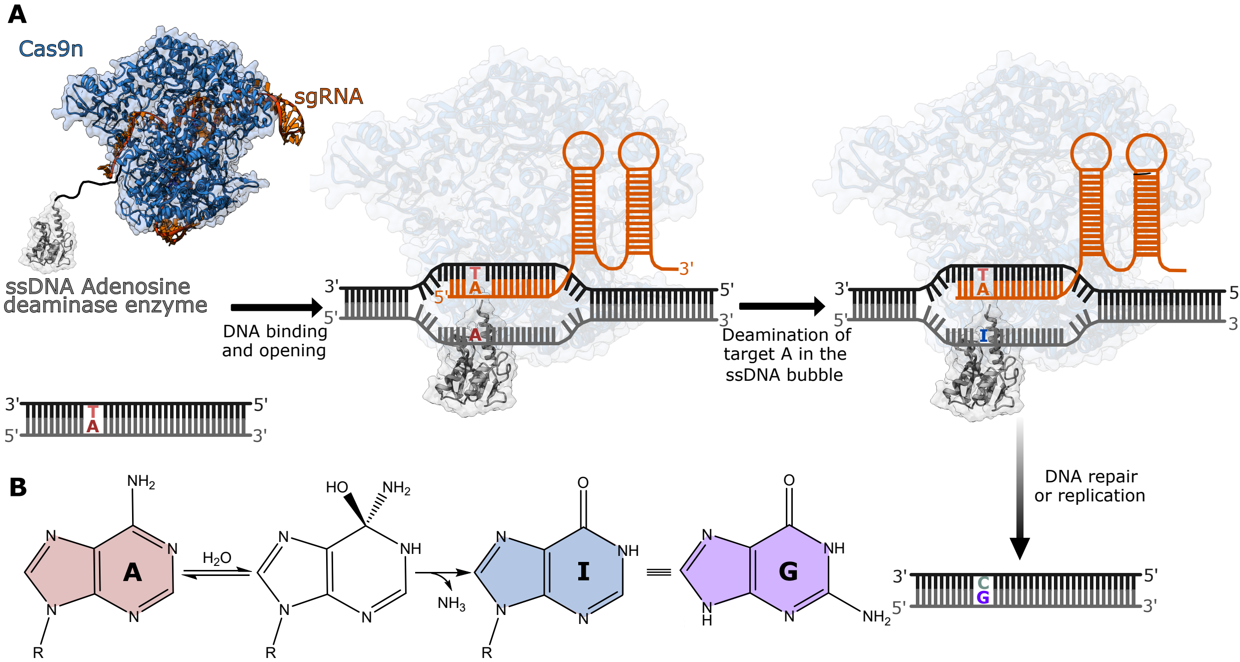

A
Cas9n
sgRNA
ssDNA Adenosine deaminase enzyme
DNA binding and opening
Deamination of target A in the ssDNA bubble
DNA repair or replication
3'
5'
T
A
I
C
G
B
NH2
H2O
HO
NH
NH3
O
N
R
A
I
G

Figure 2.1: **Mechanism of base editing by ABEs.** (a) A schematic representation of base editing by ABEs. The ABEs studied as a part of the current work consist of a Cas9n fused to an evolved TadA* protein. The binding of Cas9n to the target genomic locus unwinds the DNA double helix and exposes a small region of ssDNA. TadA* acts on this ssDNA and deaminates adenine (A) to form inosine (I), which is subsequently converted to guanine (G) through DNA repair and replication. (b) Overall chemical reaction catalyzed by ABEs.

in efficiency as compared to the dimer construct (**Figure 2.2(a)**). These results suggest that the successive rounds of evolution performed on TadA have caused the enzyme to modify ssDNA as a monomer. Therefore, the TadA monomer is the most relevant model system with which to study the enzyme in the context of its interaction with ssDNA. Wild-type TadA consists of a five-stranded β sheet core, with five α helices wrapped around to form the active site. In addition, TadA displays a long-disordered loop (24 amino acids, residue numbers 118 to 142) that joins the β4 and β5 strands (**Figure 2.2(b) and (c)**).[17] We performed 500ns all-atom MD simulations starting with the crystal structure of wild-type E. coli TadA[17] (TadA*0.1) to gain insights into the structural dynamics of the protein (see Methods). The simulations confirmed the highly fluxional nature of the β4-β5 loop in the wild-type enzyme (**Figure 2.2(b) and (c)**). To observe the effects associated with the mutations on the structure and dynamics of TadA, we subjected the TadA*0.1

model to sequential mutations at residues 108, 106, and 147 and 155 to yield the TadA*1.1, TadA*1.2, and TadA*2.1 mutants, respectively. MD simulations of the four TadA* mutants reveal that the most substantial structural difference between TadA*0.1 and the higher-generation TadA*s occurs in this β4-β5 loop. While TadA*0.1 displays high flexibility in this region, the first mutation (Asp108Asn) leads to restricted structural mobility of the loop, with the TadA*1.2 and TadA*2.1 following this same trend (**Figure 2.2(b)**). The ubiquitous nature of this change is indicated by the reduced flexibility being observed for TadA*7.10, which harbors all the 14 mutation reported in the most evolved ABE protein (Figure S1A).[11] The suppression of the loop dynamics indicates that the replacement of Asp with Asn at residue number 108 of the protein is accompanied by a gain of structure. To quantify this effect in each TadA* mutant, we clustered all the conformations sampled by the β4-β5 loop throughout the simulations into 10 structural groups representative of the conformational space. Comparison of these representative clusters reveals high variability among the loop conformations sampled by TadA*0.1 [average root mean square deviation (RMSD) = 1.75 Å; table S1], while TadA*1.1 and higher display significantly smaller differences in the orientation of the β4-β5 loop across the 10 representative structural groups (average RMSD = 0.74, 0.88, and 0.624 Å for TadA*1.1, TadA*1.2, and TadA*2.1, respectively; table S1). Our simulations also indicate that this decrease in the structural flexibility of the β4-β5 loop of the TadA* mutants (**Figure 2.2**) may be responsible for TadA* acting as a monomer to modify DNA, as it resembles the dynamics of the wtTadA dimer (Figure S2).

2.3.2 Interaction of TadA*s with ssDNA

Next, we sought to understand the functional significance of the ABE mutations in the context of ssDNA binding. The lack of any reported structure of the entire ABE-DNA complex in the literature precludes the use of MD simulations on the entire ABE complex. Since the system of interest is only the evolving monomeric TadA enzyme and its ssDNA target and the TadA*-Cas9n complex has a size of more than 200 kDa, we reduced our molecular model to a series of TadA*

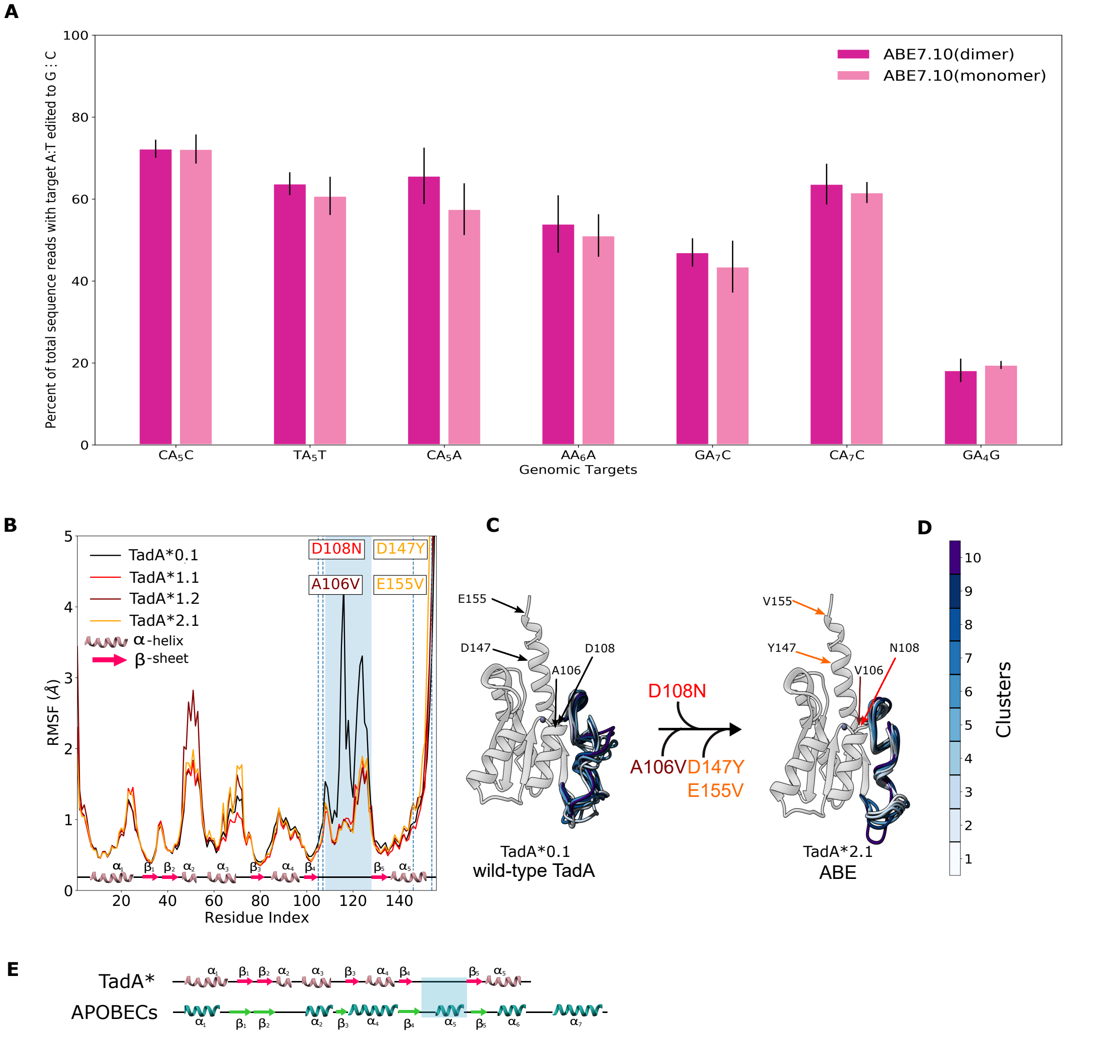


Figure 2.2: **Structural changes in the TadA* mutants revealed through MD simulations.** (a) The A•T→G•C base editing efficiency of the monomeric and dimeric ABE7.10 at six different target As in HEK293T cells. Values and error bars reflect the mean and SD of three independent biological replicates performed on different days. (b) Residue level flexibility of TadA* shown in terms of the root mean squared fluctuation (RMSF) of the Cα atoms of the peptide backbone. The β4-β5 loop region is highlighted in blue, and each mutation is indicated with its respective location in the protein. (c) Representative clusters from the trajectory of TadA*0.1and TadA*2.1 superimposed on each other, with clusters color coded as indicated in (d). (e) Comparison of the secondary structure of zinc-dependent deaminases: TadA* and APOBECs. Helices and arrows denote the α helices and β strands, respectively. The β4-β5 loop of interest in this study that interacts with the polynucleotide substrate is highlighted in both cases.

mutants in complex with a 11-mer piece of ssDNA (5′-GACT**A**CAGACT-3′). In lieu of including Cas9n and the full R-loop portions of the ABE complex, we have imposed constraints on the 5- and 3-terminal nucleotides of the ssDNA, keeping them 40 *Å* apart [based on Protein Data Bank (PDB) ID: 5y36[25]] to maintain its R-loop conformation throughout the entirety of the simulations (Methods). We then carried out unbiased MD simulations in which we allowed each of the four TadA* mutants to interact with the constrained ssDNA for 500ns and looked for changes in interactions between individual TadA* residues and the nucleic acid substrate among the four mutants. Experimentally, TadA*0.1 is not competent for base editing, but the three mutants (TadA*1.1, TadA*1.2, and TadA*2.1) are. We therefore specifically focused on identifying the interactions present in only TadA*1.1 and higher, with a particular emphasis on residue 108 (Asp in TadA*0.1 and Asn in all others), as this residue is responsible for imparting the enzyme with detectable base editing activity. To gain insights into the spatial extent of the interactions at play in the binding process, we projected the interactions between the target adenosine and its 5′- and 3′-adjacent bases (T**A**C) and the surrounding amino acids onto asteroid diagrams (**Figure 2.3(a)to (d)**). In these diagrams, we use a network representation in which these three nucleotides of the DNA are depicted as the central node and the TadA* residues are the peripheral nodes. As the typical donor atom–donor hydrogen acceptor atom distance is approximated to be 3.5 Å in globular proteins,[26, 27] we defined the first interaction shell around the DNA as all amino acids within 4 Å of the three bases in the active site. The size of each node is proportional to the time individual residues spend within the 4 Å shell during the simulation. Hydrogen bonds between residues [defined as in the CPPTRAJ package[28, 29]] are depicted as arrows connecting the corresponding nodes, with the arrow size being proportional to the hydrogen-bond strength, which is defined as the number of times that the specific hydrogen bond is established (**Figure 2.3(a)to (d)** and Figure S3). In the crystal structure of wild-type TadA in complex with its tRNA substrate [PDB ID: 2b3j[30]], Asp108 makes a hydrogen bond with the 2′-OH group of the 5′ flanking base. In contrast, when complexed with ssDNA, which lacks this hydrogen-bond donor, the repulsive electrostatic interactions between the negatively charged Asp108 and the phosphate backbone of the DNA favors a conformation in which Asp108 points toward the active site zinc ion (**Figure 2.3(e)**). Mutating Asp108 to Asn neutralizes this repulsive interaction and causes the residue to flip into a more energetically favorable conformation in which it faces the DNA substrate and interacts with the base 5 to the target adenosine. This conformational change allows Asn108 to form a hydrogen bond with the carbonyl at position 2 of the $5'$ nucleobase when this base is a pyrimidine (**Figure 2.3(f)**). This interaction between Asn108 and the $5'$ pyrimidine may explain the earlier generation ABE's strict sequence preference for a pyrimidine at this position. As subsequent mutations are introduced into TadA*, this hydrogen bond is progressively strengthened, and in the TadA*2.1 mutant, a second hydrogen bond forms between Asn108 and the phosphate backbone (**Figure 2.3(f)**. We attribute this conformational switch to the hydrogen-bond donor nature of Asn as opposed to the hydrogen-bond acceptor nature of the negatively charged Asp. The Asp147Tyr and Glu155Val mutations, which are introduced as TadA*1.2 becomes TadA*2.1, do not lie within the first interaction shell, but rather cause structural rearrangements to the protein that strengthen the interactions between Lys110, Phe148, and Phe149 and the ssDNA and cause Arg153 to become a double donor (**Figure 2.3(d) and (f), and 2.4(a)**).

2.3.3 Analyses of mutations in the α5 helix

To better understand the effects of the second-generation mutations (Asp147Tyr and Glu155Val), which are located outside of the 4 Å primary interaction shell, we expanded our analysis of the TadA*- ssDNA simulations to include the secondary interaction shell, which encompasses all residues within 4 Å of the primary interaction shell residues. Analogous to Figure 2.3, individual residues are represented by nodes whose sizes are proportional to the number of frames in the MD trajectory in which the residue lies within the specific shell, with hydrogen bonds between residues depicted as arrows between the interacting nodes, and the

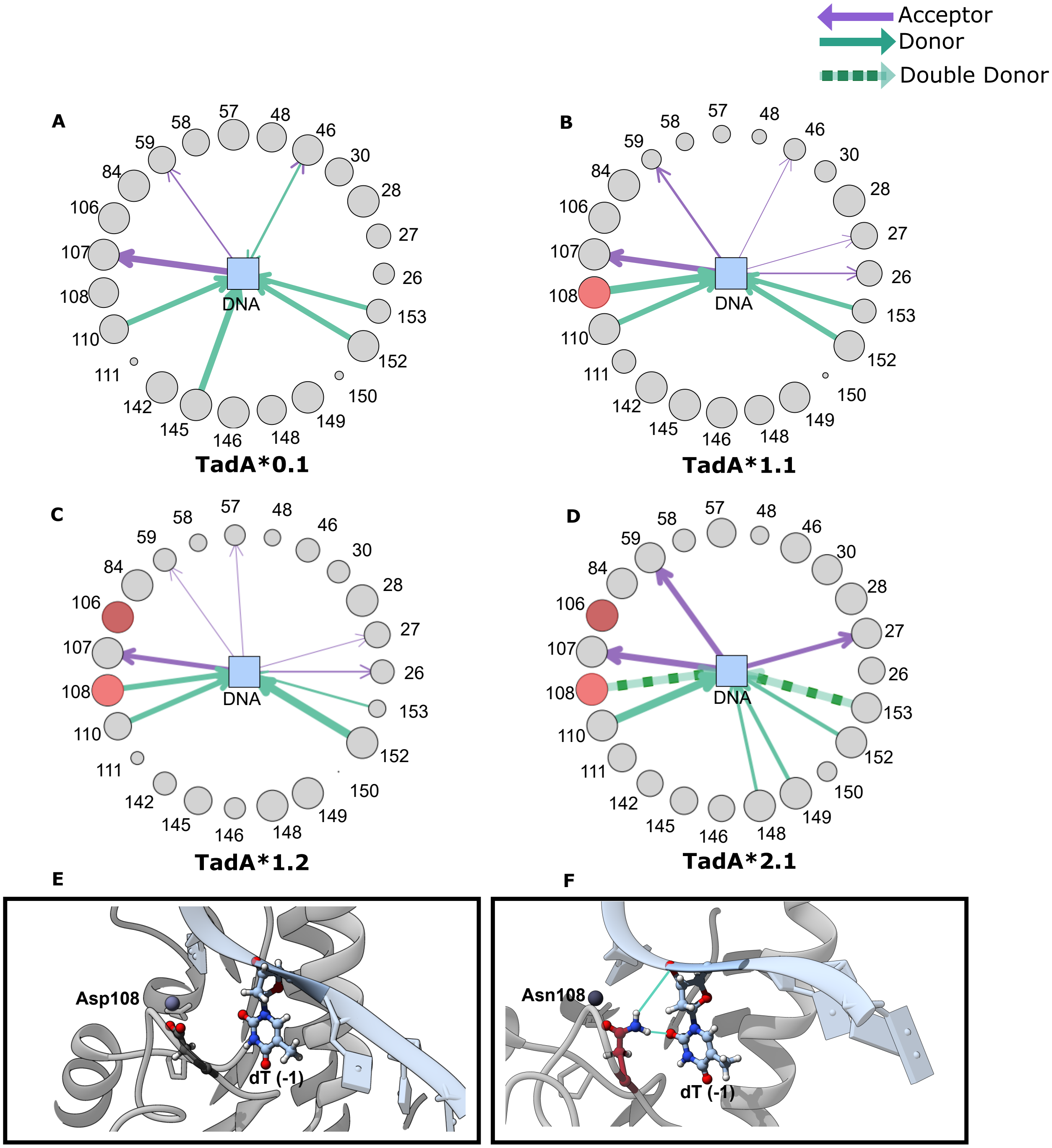


Figure 2.3: **Analyses of the TadA*-DNA contacts.** Asteroid plots for (a) TadA*0.1-ssDNA, (b) TadA*1.1-ssDNA, (c) TadA*1.2-ssDNA, and (d) TadA*2.1-ssDNA complexes. Details of the conformational change of residue 108 when it is mutated from Asp (TadA*0.1) (e) to Asn (TadA*1.1and later) (f).

arrow size being proportional to the hydrogen-bond strength (**Figure 2.4(a)**). We found that while Asp147Tyr and Glu155Val do not belong to the primary interaction shell, they do influence the manner in which the primary shell residues interact with the ssDNA. Mutation of Asp147 to Tyr abrogates a salt bridge between itself and Arg150 (primary interaction shell) that exists in TadA*0.1 (**Figure 2.2(a)**). This lost interaction results in the movement of the entire α5 helix toward the active site (**Figure 2.4(b)**), causing residues 150 to 153 to considerably spend more time within the primary interaction shell and increasing the strength of the hydrogen bonds between residues 148, 149, and 153 and the ssDNA (**Figure 2.3(a) and (d) and 2.4(a)**, and Figure S4A). Moreover, the Asp147Tyr and Glu155Val mutations, which convert negatively charged residues into neutral amino acids in the α5 helix, increase the positive charge density on the surface of the TadA*2.1 (Figure S4B and C), potentially enhancing the electrostatic interactions of the TadA* with the negatively charged ssDNA.

2.3.4 Differential binding of TadA*s to ssDNA

After qualitatively observing the interactions between the TadA* residues and the ssDNA, we sought to quantify the thermodynamics of ssDNA binding by the four TadA* mutants. To this end, we performed umbrella sampling simulations to determine the potential of mean force (PMF) associated with the binding process. In this analysis, the PMF is calculated as a function of the relative distance between the centers of mass of the ssDNA and the TadA* mutants (ξ, collective variable), which we vary from 10 to 30 Å (**Figure 2.5(a) and (b)**). The PMF profile describing the binding of TadA*0.1 to ssDNA has a minimum at ξ = 20 Å and shows a relatively small (17 kcal/mol) dissociation energy as the ssDNA is moved away from the protein to ξ = 30 Å. Once the Asp108Asn mutation in TadA*1.1 has been introduced, the PMF minimum slightly shifts toward the active site (to ξ = 18 Å), and we observe the free energy of binding increase to 42 kcal/mol as ξ is increased to 30 Å (Figure 2.5(c)). The PMF profiles calculated for the binding of TadA*1.2 and TadA*2.1 to ssDNA maintain this increased slope for ξ larger than 20 Å, implying that the

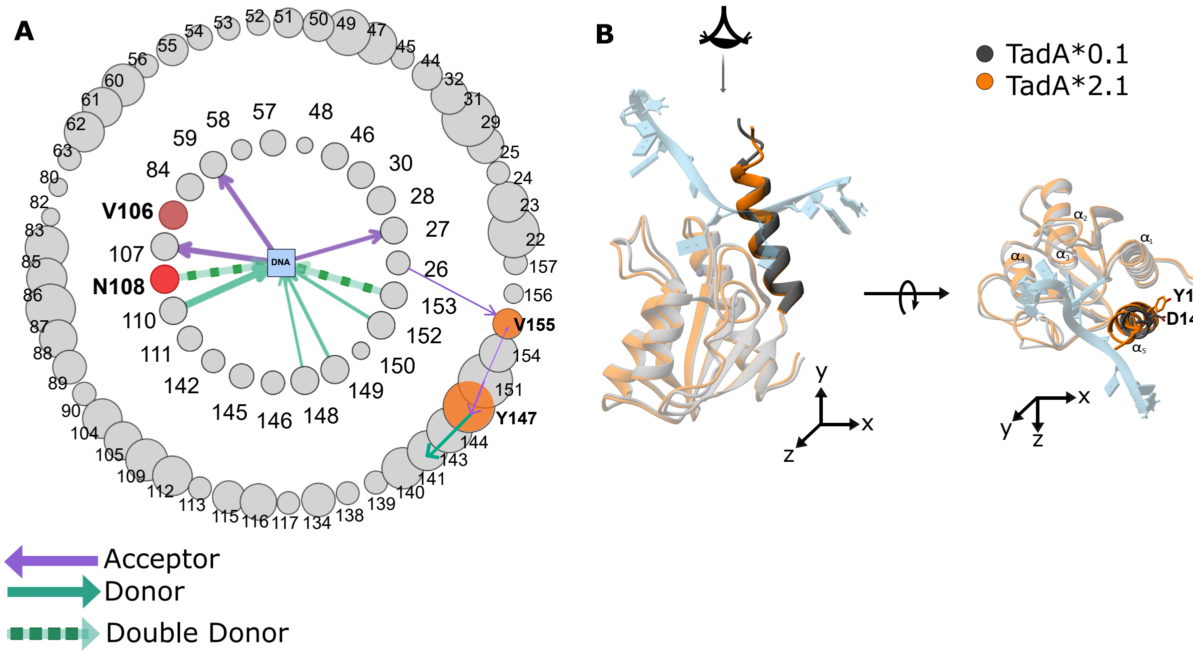


Figure 2.4: **Asteroid plot analysis of second-generation mutations.** (a) The first and second interaction shell around the three nucleotides in the active site of the TadA*2.1-ssDNA complex. The size of the node corresponds to the time in which the amino acid resides in the first/second shell. First round mutations are red, and second round mutations are orange. (b) Structural overlay of average structure of TadA*0.1-ssDNA and TadA*2.1-ssDNA complexes. This α5 helix has been highlighted to depict its overall movement toward the active site upon Asp147Tyr mutation.

single Asn108 mutation is effectively responsible for increasing the binding free energy by ≈20 kcal/mol. For ξ values less than 20 Å, the PMF profiles become sequentially more repulsive with subsequent generations, demonstrating a tighter binding of the ssDNA to the TadA*. We repeated the binding free energy calculations with a different sequence of ssDNA that lacks 5′-pyrimidine (5-GTCA**A**GAAAC-3) and again observed mutation-dependent TadA*-ssDNA binding but to a lesser extent of only 10 kcal/mol for this substrate (Figure S5). These results are in agreement with experimental observations that these early generation ABE mutants had a strong preference for YAC (Y = pyrimidine) sequence motifs. These findings highlight the importance of the Asp108Asn mutation in imparting functional promiscuity to the TadA* enzyme toward ssDNA editing[11] through an increase in the free energy of binding. While the binding affinity is not a

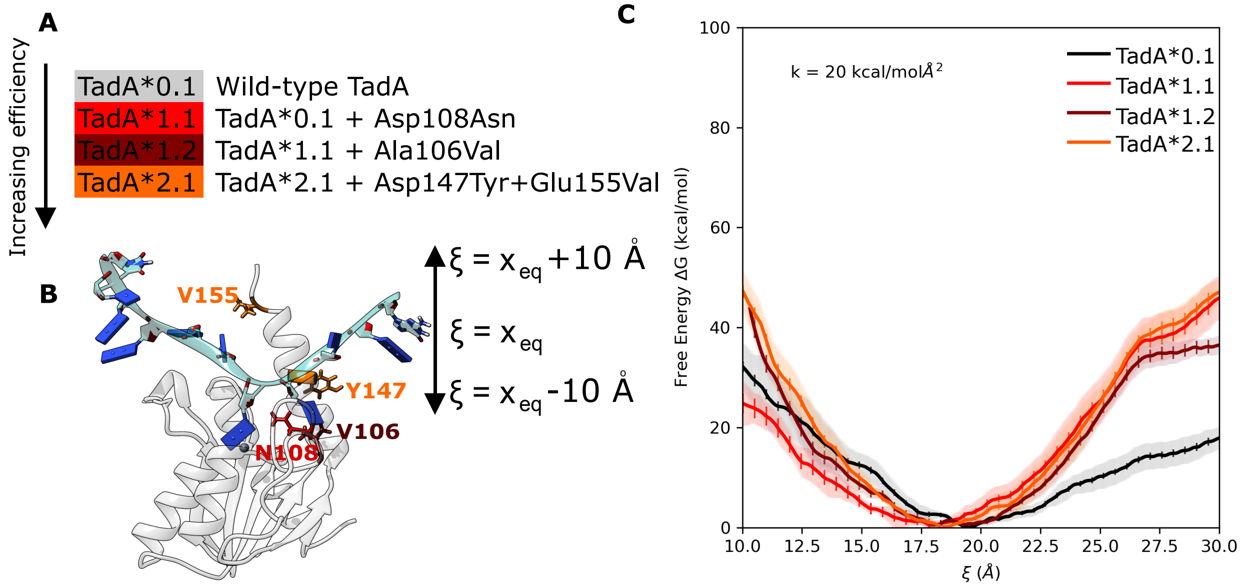


Figure 2.5: **ABE mutations lead to an increase in TadA* binding to the ssDNA.** (a) List of early generation mutations in TadA that were analyzed in this study. (b) The model of the TadA*-ssDNA complex simulated to determine the binding energy profile of the TadA* mutants. The binding-unbinding event was monitored using the collective variable (ξ) defined as the distance between the center of mass of the protein and DNA. (c) The free-energy profile of binding of the ssDNA to various TadA*s. For each TadA*-ssDNA complex, the average PMF is shown as a function of the continuously changing ξ values. The shaded regions around individual curves depict the standard deviation for four independent replicates of the umbrella sampling simulations. The error bars associated with the mean PMFs indicate the error calculated using the block-averaging method.

direct measure of the editing efficiency, our analyses of the TadA*-ssDNA complexes demonstrate that the initial Asp108Asn mutation, which plays a critical role in the onset of the DNA editing capability of the ABEs, leads to increased binding between the TadA* and the ssDNA substrate. We speculate that higher-generation mutations take advantage of this increased binding to improve the kinetics of base editing and broaden the substrate sequence scope.

2.3.5 Reversion analysis of Asn108 mutation

To confirm the crucial role played by Asn108 in ssDNA editing by ABE, we subjected the higher generation of TadA* mutants (TadA*1.2 and TadA*2.1) to reversion analysis of this mutation. Specifically, by mutating Asn108 back to Asp108 in both TadA*1.2 and TadA*2.1, we generated two new TadA mutants, TadA*1.2(N108D) and TadA*2.1(N108D), respectively (**Figure 2.6(a)**).

To disentangle the structural contribution of Ala106Val, Asp147Tyr, and Glu155Val from that of Asp108Asn, we monitored the structural flexibility of TadA*1.2(N108D) and TadA*2.1(N108D) (**Figure 2.6(b)**). We observed the maintenance of the β4-β5 loop stabilization, suggesting that the Ala106Val mutation is also sufficient to induce this change in structural flexibility (Figure S6). We also observed a slight increase in the flexibility of the α2 helix due to this mutation, but upon introduction of the round two mutations, this is lost. To complement these structural studies, we also characterized the binding free energy in the TadA*1.2(N108D)-ssDNA and TadA*2.1(N108D)- ssDNA complexes. Unlike the structural results and despite having respectively one and three mutations that were experimentally found to be favorable for ssDNA editing, TadA*1.2(N108D) and TadA*2.1(N108D) produced PMF profiles that are significantly different from those of their parent mutants (**Figure 2.6(c)**). In particular, both PMFs closely follow the corresponding profile obtained for TadA*0.1 for ξ values larger than 20 Å yet are considerably more repulsive for ξ values smaller than 20 Å. We performed analogous reversion analysis for the TadA*7.10 (which contains all 14 identified mutations) and observed qualitatively similar trends for the TadA*7.10(N108D) (Figure S7A).

These differences demonstrate weaker binding between the ssDNA and ABE mutants lacking the Asn108 mutation. To confirm our computational results, we generated the ABE1.2(N108D) and ABE2.1(N108D) constructs and experimentally measured their respective A•T→G•C base editing efficiencies using high-throughput sequencing (HTS) alongside ABE0.1, ABE1.2, and ABE2.1 in HEK293T cells at six different targets. Reversion of Asn108 mutation to Asp led to an average decrease in the A•T→G•C base editing efficiency of 22-fold (ranging from 6.5- to 42-fold) and 70-fold (ranging from 22.6- to 126-fold) for ABE1.2 and ABE2.1, respectively (Figure 2.6(d)). It is notable that even the presence of all three Ala106Val, Asp147Tyr, and

Glu155Val mutations was not sufficient to restore editing activity with Asp at position 108; both ABE1.2(N108D) and ABE2.1 (N108D) induced average A•T→G•C base editing efficiencies of 0.36 and 0.29% across all six editable As, as compared to 3.6 and 16.8% for their respective parental mutants. Reversion of the Asn108 mutation in the ABE7.10 background displayed a similar trend. Replacement of Asn108 with Asp in both monomeric and dimeric ABE7.10 decreased the A•T→G•C base editing efficiency by an average factor of 146-fold (ranging from 67- to 176-fold) and 123-fold (ranging from 35-fold to 259-fold), respectively (**Figure 2.6(e)**). This indicates that the presence of 13 higher generation mutations, independently of being installed in the monomeric or dimeric construct, cannot compensate for the loss of the Asn108 mutation. The importance of residue Asn108 in ABE7.10 was also recognized in the experimental study by Rees *et al.*,[31] where radical substitutions of Asn108 with Phe, Trp, and Met were found to result in complete abolishment of any DNA editing activity at all target adenosines except when the target nucleobase was at position 5 within the protospacer. However, conservative substitutions of Asn108 with Gln, and Lys, resulted in decreased DNA editing efficiencies for these mutants, albeit in a sequence-dependent manner and to a much smaller extent than the substitution with Asp.[31] The results of this study thus provide further support of the hydrogen-bonding analysis presented here, which emphasizes the requirement of a positive charge density, either in the form of a hydrogen-bond donor as Asn (**Figure 2.4**) or Gln[31] or a positively charged residue as Lys[31] for enabling the ssDNA activity of TadA*. Collectively, these data demonstrate the drastic effects a single atom substitution (from N to O) can have on protein function and highlight the complexity of protein sequence- structure-function relationships.

2.4 Discussion

Enhancing our understanding of how an enzyme's sequence influences its function will help increase the success of future directed evolution projects. Although the mutations discovered

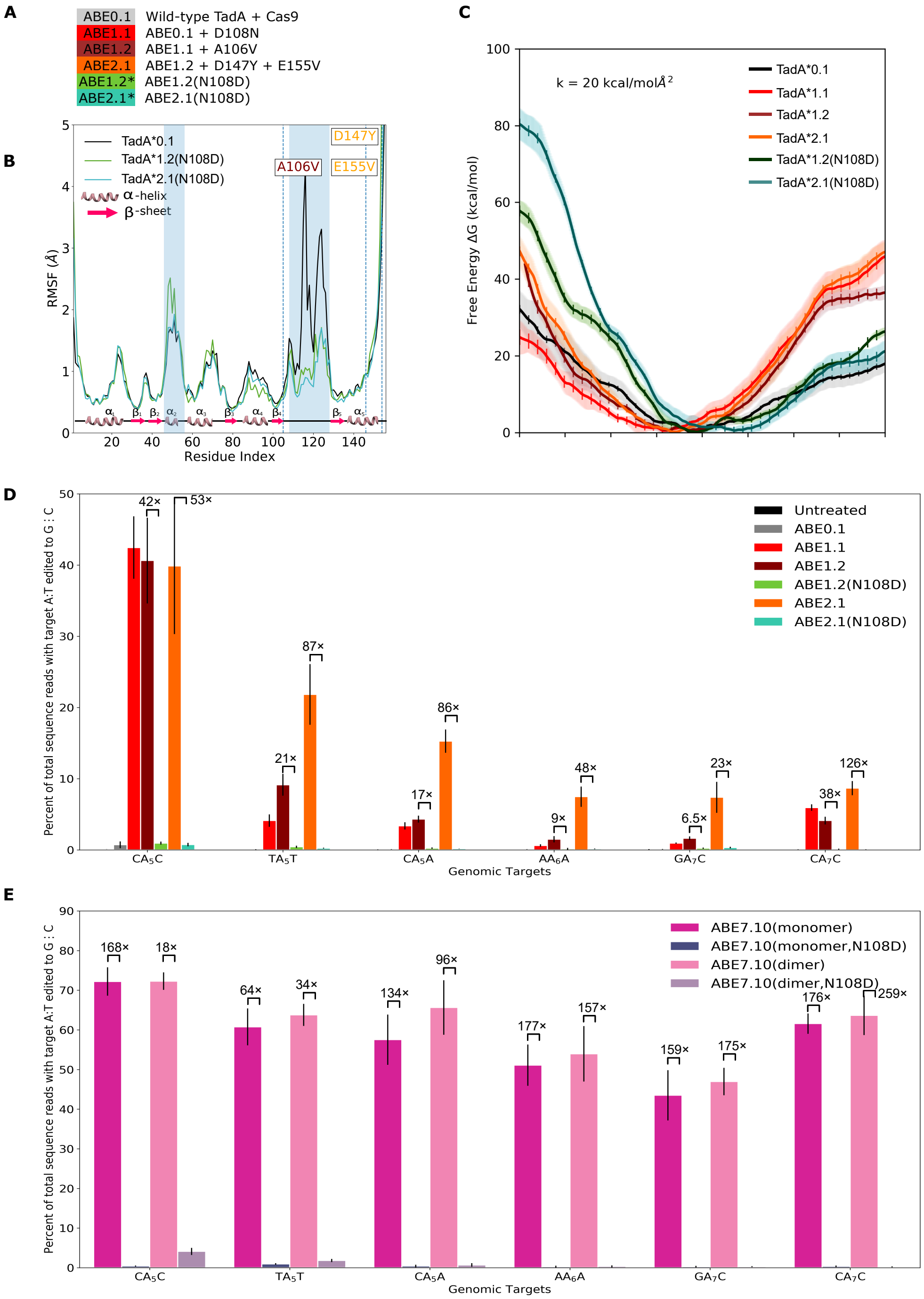


Figure 2.6: **Significance of Asn108 for base editing. (a) ABE constructs created by reverting the Asp108Asn mutation in the higher generation ABEs.** (b) RMSF of the Cα atoms of the TadA*1.2(N108D) and TadA*2.1(N108D) enzymes. (c) The free-energy profile of binding of the hybrid TadA*s to ssDNA. The shaded regions around individual curves depict the SD for four independent replicates of the umbrella sampling calculations. The error bars associated with the mean PMFs indicate the error calculated using block-averaging method. (d and e) A•T→G•C base editing efficiencies in HEK293T cells by the various ABEs at six different target As. Fold-decrease values upon reversion analysis of the Asp108Asn mutation are indicated above the bars. Values and error bars reflect the mean and SD of three independent biological replicates performed on different days.

using directed evolution are exceptional at enhancing the particular enzymatic property being pursued, these mutations are difficult to predict and require considerable experimental resources. As the development of future base editors will likely involve additional directed evolution efforts,[32,33] maximizing our understanding of the outcomes of previous studies on this front will aid in these future studies. This work is an *a posteriori* study using a combination of computational simulations and experimental measurements to understand the mutations generated during the directed evolution of ABEs.[11] We have additionally carried out MD simulations of TadA* and TadA*-ssDNA models to explore how the initial mutations accumulated during directed evolution give rise to ssDNA editing by the ABE enzyme. Installation of the Asp108Asn mutation in the TadA*0.1 to generate TadA*1.1 leads to a significant decrease in the flexibility of the β4-β5 loop of the TadA (**Figure 2.2**). This loop is known to both impart sequence specificity to the wild-type TadA enzyme through interactions with the nucleobases immediately upstream of the target A base and also serve as the dimerization interface between the individual TadA proteins.[30] Our simulations indicate that the structural dynamics of TadA* mutants (**Figure 2.2**) resembles that of the wtTadA dimer (Figure S2), which may explain how the TadA* enzymes are performing DNA base editing as monomers. The changes observed in the dynamics of the β4-β5 loop therefore may help broaden the substrate scope of the TadA* enzymes to include both tRNA and ssDNA. In addition, as the TadA* mutants were evolved to function as monomers, this change in the dynamics may be increasing the enzyme's affinity for ssDNA at the expense of protein dimerization. This is proven to be the case, as we experimentally observe that the TadA enzyme works as a monomer when acting on ssDNA, a finding that represents a key step in characterizing the mechanism of base editing by ABE (**Figure 2.2**). This is an unexpected result that fundamentally changes our understanding of how ABEs function and will likely affect future ABE engineering and optimization studies. Intriguingly, loss of conformational flexibility in the β4-β5 loop of TadA* appears to make the overall structure of the protein more analogous to the APOBEC family of proteins (**Figure 2.2(e)**). APOBEC enzymes are a class of proteins that have cytidine deamination activity on both ssDNA and ssRNA[20,21] and were repurposed into the original CBEs. The inherent nature of the APOBECs to edit a broad range of nucleotide targets is preserved in the CBEs, which have been shown to exhibit considerable off-target DNA and RNA activities due to the APOBEC1 portion of the base editor.[34,35] This dual-substrate specificity of APOBECs has been attributed to specific conformations of the active site loop (α1-β1 loop, β2-α2 loop, and β4-β5 also referred to as the loop 1, loop 3, and loop 7, respectively) that interacts with the 5′ flanking base of the substrate nucleotide using both experiments and simulations.[36–39] Both TadA and the APOBEC enzymes share a core five-stranded β sheet structural element surrounded by α helices. The β4-β5 loop serves the same functional purpose in both enzymes, but the length of this loop is substantially longer in TadA, and in the APOBECs, it assumes a definite α-helical secondary structure (**Figure 2.2(e)** and Figure S8). The gain in structure of this loop in TadA may contribute to the gain of ssDNA editing capability by TadA*,[20,40] but it is not solely responsible for this activity. The TadA*1.2(N108D) enzyme retains reduced mobility in the β4-β5 loop yet displays wild-type like ssDNA binding affinity according to our simulations and nearly undetectable base editing efficiencies in our experimental work. Note that the Ala106Val mutation causes a substantial gain in mobility of the α2 helix (**Figure 2.2(a) and 2.6(b)**), which is canceled out when the Asp147Tyr and Glu155Val mutations are incorporated. The α2 helix of TadA aligns with the β2-α2 active site loop of the APOBECs (**Figure 2.2(e)** and Figure . S8), which lacks secondary structure and has been shown to be responsible for sequence specificity of the enzymes. Our simulations show that when wild-type TadA interacts with ssDNA, the absence of a hydrogen-bond donor (in the form of the 2′-OH group of the ribose sugar in RNA) for Asp108 causes this residue to flip into an energetically unfavorable conformation away from the negatively charged DNA backbone. This unfavorable conformation is responsible for the lack of ssDNA editing by the wild-type enzyme, as the presence of all other 13 favorable mutations, and the favorable interactions they bring with them, is not enough to compensate for the strained configuration that Asp108 is forced to adopt when in the presence of DNA rather than RNA. However, upon neutralization of this negative charge when Asp108 is mutated to Asn (a single atom substitution from O to N), the residue can now rotate back into a more energetically favorable position, allowing for the enzyme to interact with ssDNA. This rotation toward the ssDNA substrate also allows for the formation of a hydrogen bond between residue 108 in TadA* and the ssDNA (the -1 nucleotide in **Figure 2.3(d) and (e)**. This hydrogen bond further strengthened in TadA*2.1, where Asn108 becomes a double hydrogen-bond donor, interacting with the phosphate backbone. The phosphate backbone is a structural element common to both DNA and RNA, suggesting that in the process of acquiring ssDNA editing capabilities, TadA* may not surrender its native RNA editing functionality. This has been confirmed by previous reports of off-target RNA editing by ABE enzymes.[31,35] Furthermore, it was recently found that removing wtTadA from ABE7.10 does not suppress its RNA deamination activity, which demonstrates that the Asp108Asn mutation supports RNA binding by TadA*.[41] While one may expect only residues in the first shell (that interact directly with the ssDNA) to be primarily responsible for enhancing the thermodynamics and kinetics of ssDNA editing by TadA*, 6 of the 14 overall mutations accumulated during directed evolution actually reside in the second shell of the enzyme (Figure S1). In addition to electrostatic contributions, through our simulations, we observed that the Asp147Tyr and Glu155Val mutations, both of which reside in the α5 helix (Figure S4B), cause structural rearrangements in the protein, effectively initiating a chain reaction that strengthens the interactions between a variety of primary shell residues and the ssDNA substrate. Note that nearly half (6 of 14) of the ABE7.10 mutations are located in the α5 helix, highlighting the significance of understanding its role in ssDNA editing. These enhanced hydrogen-bonding interaction between the TadA* residues and the ssDNA, caused in aggregate by all four mutations, and the now-favorable conformation of the residue 108 when it is Asn, also translate into an increased free energy binding of the TadA*s to ssDNA (**Figure 2.5** and Figure S5). Upon reversion of Asn108 to Asp, however, even in the presence of the three other advantageous mutations (Ala106Val, Asp147Tyr, and Glu155Val), we observe a marked decrease in the binding affinity of TadA*1.2(N108D) and TadA*2.1(N108D) to ssDNA (**Figure 2.6(c)**). On the basis of these observations, we speculate that the Asp108Asn mutation may play a bipartite role: It affords structural rigidity to the region of the enzyme responsible for sequence specificity and increases the binding affinity of the TadA enzyme to ssDNA through hydrogen-bonding interactions. However, the hydrogen bonds that Asn108 forms with the 5′ nucleobase and the phosphate backbone are not its only contribution to the onset of DNA editing activity by ABEs. Simulations and experiments verify that reversion of Asn108 back to Asp from higher-generation ABEs leads to nearly complete loss in the base editing activities of higher ABE mutants (**Figure 2.6**), despite the presence of up to 13 other beneficial mutations in TadA* that have created additional hydrogen-bonding interactions between TadA* and the ssDNA (Figure 2.3(d) and 2.4). It is likely that the increased conformational strain imposed on the Asp108 residue when it must flip around to point away from the DNA backbone is energetically unfavorable enough to preclude ssDNA binding even with these additional favorable hydrogen-bonding interactions. This study provides the first insights into the mechanism of base editing by ABEs, beginning with the observation that the TadA* enzyme acts a monomer to modify ssDNA. The results presented in this study additionally provide an explanation of the structural and functional roles of the initial TadA mutations identified in the evolution of ABE. We anticipate that this atomistic understanding of previous successful directed evolution experiments will enable the prediction of new mutations and lead to the rational engineering of future base editors.

2.5 Methods

The crystal structure of *E. coli* TadA enzyme (PDB ID: 1z3a) was used to define the initial coordinates for TadA*0.1.[17] The TadA*1.1, TadA*1.2, TadA*2.1, TadA*1.2(N108D), and TadA*2.1(N108D) mutants were prepared by inducing virtual mutations to the TadA0.1 structure using the mutagenesis plugin available in PyMOL. We then combined the crystal structure of E.

coli TadA enzyme with the tRNA substrate from its structural homolog from Staphylococcus aureus [PDB ID: 2b3j[30]] to prepare the TadA*-ssDNA complexes. The remodeling of the tRNA structure by the removal of the $2'$ hydroxyl groups and all changes in the sugar pucker of the nucleotide backbone were carried out using the swapna command in the Chimera software.[42] Moreover, since the tRNA structure was crystallized bearing nebularine, a nonhydrolyzable adenosine analog (18), we used the swapna command to substitute nebularine with adenine. To unpair the $3'$ and $5'$ ends of the hairpin loop, we used steered MD simulations using the exposed ssDNA nucleotides of the ternary complex of the cryo–electron microscopy structure of CRISPR-Cas9 [PDB ID: 5y36 (13)] as a reference structure (Figure . S9). This yielded the TadA*0.1-ssDNA complex as illustrated in Figure . S9. Similarly, the complexes of TadA*1.1, TadA*1.2, TadA*2.1, TadA*1.2(N108D), and TadA*2.1(N108D) mutants with ssDNA were developed using the mutagenesis plugin of PyMOL. All crystallographic water molecules within 3 Å distance of the protein/protein-ssDNA surface were preserved during the modeling process, and each of the systems was solvated using a truncated octahedral box of TIP3P water molecules.[42] All titratable residues were assigned protonation states at pH 7 as predicted by the H++ server.[43,44] Varying number of Na+ ions were added to each system to maintain charge neutrality. The protein and the DNA atoms were represented using the Amber ff14SB force field and the bsc1 parameters, respectively.[45–47] All MD simulations were performed under periodic boundary conditions using the CUDA accelerated version of PMEMD implemented in Amber18 suite of programs.[48–50] The structures were first relaxed using a combination of steepest descent and conjugate gradient minimization. This was followed by a 1-ns heating to 298.15K and 10-ns equilibration under harmonic restraints. Subsequently, we removed all restraints (except on the $5'$ and $3'$ termini of the substrate DNA sequence) and carried out 500-ns unbiased MD simulations for the six TadA* mutants and corresponding TadA*-ssDNA complexes. Additional details of this protocol can be found in Supplementary Materials and Methods on publication associated with this chapter. Table S2 summarizes all the simulations that were carried out during this study. We calculated the free-energy binding profiles of the TadA*-ssDNA complexes along the collective variable corresponding to the distance between the centers of mass of the protein and the ssDNA substrate. For each TadA*-ssDNA complex, the PMF along this collective variable was calculated using umbrella sampling simulations. Starting from the equilibrated TadA*-ssDNA structures, we conducted four independent sets of umbrella sampling simulations for all of the six TadA*-ssDNA complexes, and the final PMFs were reconstructed using the weighted histogram analysis method (WHAM) algorithm.[51] Additional error analysis was carried out using a custom block averaging script based on the method described by Zhu and Hummer.[52] The CPPTRAJ module implemented within Amber18 was used to analyze all the MD trajectories.[28,29] The root mean squared fluctuation of the ABE mutants and clustering of configurations from each MD trajectory were calculated, with respect to the Cα atoms of the protein backbone. We identified the primary and secondary interaction shells and the associated H-bonding network using the mask and hbond keywords of CPPTRAJ, respectively (see the Supplementary Materials for details). The PDB2PQR webserver, in conjunction with the APBS server, was used to calculate the electrostatic maps for the ABE0.1 and ABE2.1 models.[53] The visualization of the MD trajectories was rendered using Chimera, and data were plotted using Matplotlib.[54]

2.6 Acknowledgments

We thank M. Norman for a Director's Discretionary Allocation on the Comet GPU cluster at the San Diego Supercomputer Center. K.L.R. thanks C. Egan and E. Lambros for helpful discussions. Funding: This research was supported by the University of California San Diego and the NIH through grant no. 1R21GM135736-01. All computer simulations used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF through grant no. ACI-1548562.

Chapter 2 is reproduced, in part, with permission, from: Rallapalli, K.L., Komor, A.C.,

Paesani, F. “Computer simulations explain mutation-induced effects on the DNA editing by adenine base editors”, *Sci. Adv.* 6, eaaz2309 (2020). The dissertation author was the primary author on all reprinted materials.

Chapter 3

Understanding the off-target RNA-editing activity of ABEs

3.1 Abstract

Adenine base editors (ABEs) have been subjected to multiple rounds of mutagenesis with the goal of optimizing their function as efficient and precise genome editing agents. Despite this ever-increasing data accumulation of the effects that these mutations have on the activity of ABEs, the molecular mechanisms defining these changes in activity remain to be elucidated. In this study, we provide a systematic interpretation of the nature of these mutations using an entropy-based classification model that relies on evolutionary data from extant protein sequences. Using this model in conjunction with experimental analyses, we identify two previously reported mutations that form an epistatic pair in the RNA-editing functional landscape of ABEs. Molecular dynamics simulations reveal the atomistic details of how these two mutations affect substrate-binding and catalytic activity, via both individual and cooperative effects, hence providing insights into the mechanisms through which these two mutations are epistatically coupled.

3.2 Introduction

The ability to introduce A•T to G•C base pair conversion in the genetic code of an organism, in a precise, efficient, and programmable manner, has the potential to correct almost 60% of known pathogenic point mutations in human beings.[5] Targeted A•T to G•C conversions have recently been realized through the development of adenine base editors (ABEs).[11] ABEs consist of two subunits: a catalytically-impaired Cas9 (Cas9n), which serves as a programmable DNA-targeting module, and an engineered variant of a tRNA adenosine deaminase enzyme (TadA*),[55] which serves as the single-stranded DNA (ssDNA)-editing module and enables the hydrolytic deamination of targeted adenosines (A) into inosines (I). Inosine is subsequently converted into guanosine (G) by the DNA repair and replication machinery, completing the A•T to G•C base pair conversion by ABEs (**Figure 3.1(A)**).

ABEs continue to remain a focal point of interest for the genome editing community, not only because of their potential as therapeutic agents[56–61] but also because of the remarkable scientific effort that went into their development. Extensive protein engineering and evolutionary methods were employed to impart ssDNA-editing capabilities onto an RNA-editing enzyme, wild-type *E. coli* TadA (wtTadA), resulting in the seminal ABE7.10 base editor.[11]

Although the mutations that gave rise to the original ABE7.10 construct successfully imparted ssDNA-editing capability onto TadA*, they did not suppress the inherent RNA-editing activity of TadA*. It was subsequently demonstrated that ABE7.10 induces considerable gRNA-independent off-target RNA-editing throughout the transcriptome.[35,62,63]

Since the development of ABE7.10, major efforts have been devoted to its further evolvement on two separate fronts (**Figure 3.1(B)**). First, additional rounds of directed evolution were employed to increase the on-target ssDNA-editing activity by TadA, resulting in ABE8.20[64] and ABE8e.[65] Second, structural analyses of the TadA–RNA complex followed by rational engineering was employed to decrease the off-target RNA-editing activity by TadA, resulting in

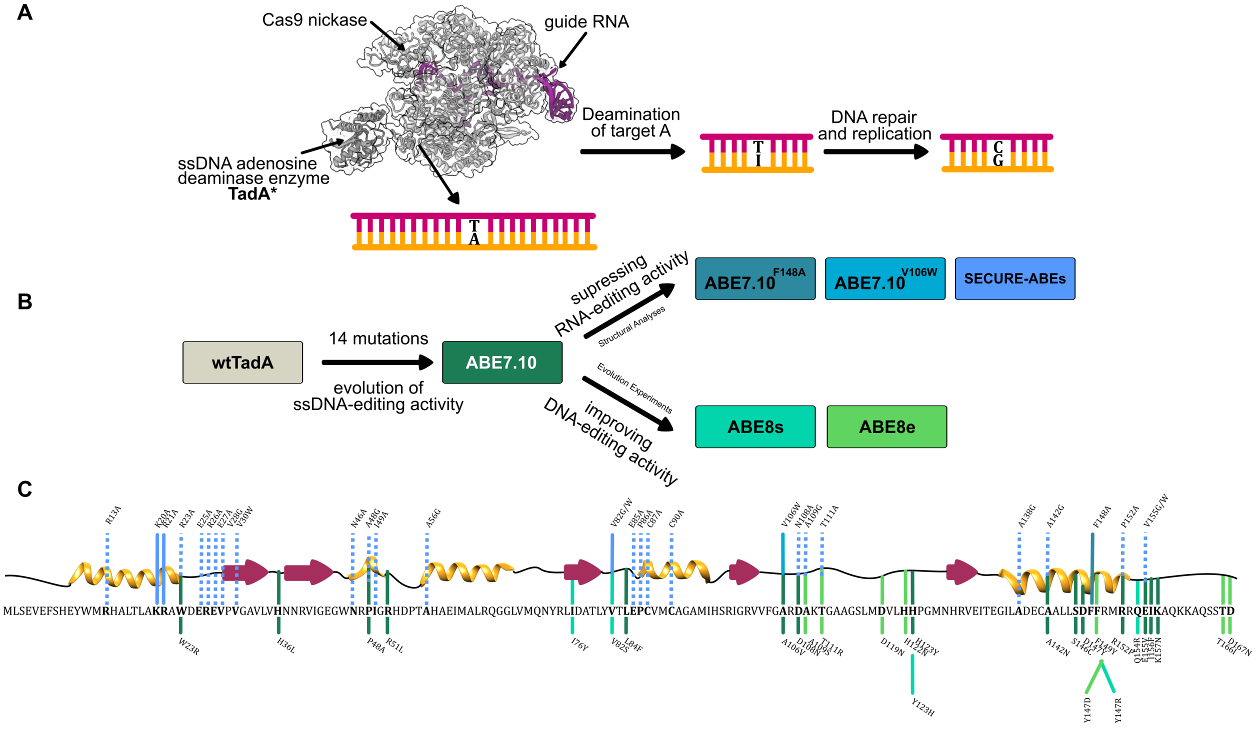


Figure 3.1: **Evolutionary Trajectory of ABEs.** (A) Schematic representation of base editing by ABEs (PDB ID: 6VPC).[67] The binding of Cas9n to the target genomic locus unwinds the DNA double helix and exposes a small region of ssDNA. TadA* hydrolytically deaminates the adenine (**A**) to form inosine (**I**) which is subsequently converted to guanine (**G**) by cellular DNA repair and replication machinery. (B) Engineering efforts in the field to generate and improve upon ABEs, starting from *E. coli* wtTadA. (C) Primary and secondary structure of *E. coli* wtTadA with key mutations indicated. The line colors correspond to colors shown in (B), indicating the ABE version in which these mutations were identified. Solid lines are mutations that were incorporated into final ABEs constructs, while dashed lines are mutations that were experimentally tested in previous work, but not incorporated into final ABE constructs.

ABE7.10^{F148A},[66] ABE7.10^{V106W},[31] and SECURE-ABEs.[41]

Due to the lack of naturally occurring ssDNA-editing enzymes (cytosine deaminases are a rare exception[10]), the expansion of the existing base editing repertoire would inevitably require evolution and engineering strategies on new enzymes to first introduce ssDNA-editing activity, followed by structure-based redesign to abrogate inherent RNA-editing activity, analogous to those used in the development of ABEs. The success of structure-based protein engineering efforts are highly dependent on the availability of appropriate X-ray or cryo-EM structures of the protein–RNA complex. Even when structural data are available to guide this process, most mutations, especially those concentrated near the active site of the enzyme, are likely to have detrimental effects on the enzyme's function.[31,41,68] Hence, it is important to fully understand the features that are essential for the native RNA-editing function of TadA*, and how certain mutations can affect changes to its substrate binding and catalytic mechanism.

To date, many studies have mutated ABEs to manipulate its DNA- and RNA-editing abilities, producing a large amount of experimental data associated with mutations at 45 of the 167 residue sites of TadA* (**Figure 3.1(B) and (C)**). To gain fundamental insights into ABE's editing activity from this ever-expanding pool of mutational information and guide future efforts in the development of new base editors, we have carried out a systematic data-driven computational study combined with experimental assays to better understand, in atomistic detail, the effects of individual mutations on the activity of TadA*.

3.3 Results

3.3.1 Dataset Curation and Sequence Entropy Calculation

The principal tenet of biochemistry is that the primary sequence of amino acids comprising a protein dictates its three-dimensional molecular structure, which then determines its biological function.[69] To date, most ABE engineering efforts have relied on the second and third tiers of this tenet, in the form of structural analyses[31,41,66] followed by site-directed mutagenesis and experimental measurements of the resulting functional properties of TadA (second tier) or directed evolution where TadA is randomly mutated and functional variants are identified through a selection scheme (third tier)[11,64,65] (**Figure 3.1(B)**). Due to the time- and resource-intensive nature of these second and third tier methods, we decided to begin our investigations by focusing instead on the first tier of this tenet, that is, on investigating how the primary sequence of TadA can be used to rationalize the effects that individual mutations, identified experimentally, have on the native function of TadA* (i.e. RNA-editing activity) (**Figure 3.1(C)**).

With the expansion of reliable protein sequence databases,[70,71] the statistical analyses of protein homologs have already enabled the successful prediction of mutational effects on the function of several enzymes,[72] including cytosine base editors.[73] For our sequence-based analyses of the ABE mutations, we used the amino acid sequence of *E. coli* wtTadA[55] as our query for a BLAST search[74] of its extant homologs in the SWISSPROT database,[71] which generated a dataset of 75 homologs. However, as our primary focus was to identify residues essential for the function of TadA on its native RNA substrate, we filtered out distant homologs using stringent percentage identity and coverage length cutoff values (**Figure 3.2(A)**). This filtering resulted in a more focused dataset as it removed functionally distinct and redundant sequences from our initial BLAST search. Despite reducing the size of the dataset considerably (to 35 homologs), this filtered dataset still captures the diversity of our initial unfiltered dataset (**Figure 3.2(A)**).

To visualize the sequence space captured by our unfiltered and filtered datasets, and highlight relationships among these wtTadA homologs, we performed a dimensionality reduction of the dataset using principal component analysis (PCA). This allows us to project the hyper-dimensional sequence space associated with the homologs onto two dimensions, while still preserving the relationships among the various homologs. By partitioning the dataset into four representative clusters (Figure S1), the outcome of filtering becomes more apparent as each cluster consists of functionally similar homologs. These clusters are represented by different colors (purple, brown, red, and blue) in **Figure 3.2(A)**. From this analysis, it can be observed that this filtered dataset indeed captures the overall diversity of the unfiltered dataset, as three of the four clusters are represented. The purple cluster (containing the query sequence) consists primarily of TadAs and their eukaryotic equivalents, ADAT2s, and hence has the greatest number of filtered sequences, 26 of the 35 filtered sequences. The next most populated cluster, the brown cluster, consists of 22 sequences in the unfiltered set, which results in 6 sequences after filtering. Only three out of eleven sequences were selected from the red cluster, with two of these sequences belonging to the cytidine deaminase superfamily (and have 50% similar to the query sequence) and the third sequence corresponding to a guanine deaminase. Given the distance between the blue cluster and the query sequence (i.e., the lack of similarity between these sequences, as they represent the catalytically inactive Tad3 and ADAT3 proteins), it is expected that no sequences were selected from this cluster upon the implementation of our filters. It is important to note that our filtered dataset consists entirely of RNA-editing enzymes, demonstrating the effectiveness of our filters. We, therefore, reasoned that any primary sequence analyses of our filtered dataset would be highly biased towards illuminating aspects of RNA-editing activity by the wtTadA (Figure S1B).

Having obtained a reliable dataset of extant TadA homologs, we next sought to quantify the evolutionary conservation and functional importance of individual residues of wtTadA. An extensively studied and widely used approach to address this problem is the evaluation of information theory-inspired sequence entropy scores (**Figure 3.2(B)**).[75–78] Within this approach, the sequence entropy for each residue site, *i*, is defined as:

$$H_i \approx -\sum_{n=1}^{N} p(i_n)\, log_{20}\, p(i_n) \quad \text{for} \quad i \in \{1,\ldots,L\} \tag{3.1}$$

where $p(i_n)$ refers to the statistical probability of having a particular amino acid n at site i and N is the total number of amino acids. Thus, the value of H_i ranges between 0 and 1, with an entropy value of 0 indicating that the site has only one unique amino acid represented within the dataset (suggesting that the site is highly conserved from an evolutionary standpoint), and an entropy value of 1 indicating that the site has every possible amino acid represented within the dataset (suggesting that such a site is naturally more tolerant to mutations).

Applying Equation 3.2 to the filtered dataset of TadA homologs (**Figure 3.2(A)**), we calculated the site entropy for the entire wtTadA sequence (**Figure 3.2(C)**). The active site of wtTadA consists of a zinc ion tetrahedrally coordinated by Cys^{87}, Cys^{90}, His^{57}, and a water molecule. This water molecule is activated for deamination reaction by the highly conserved

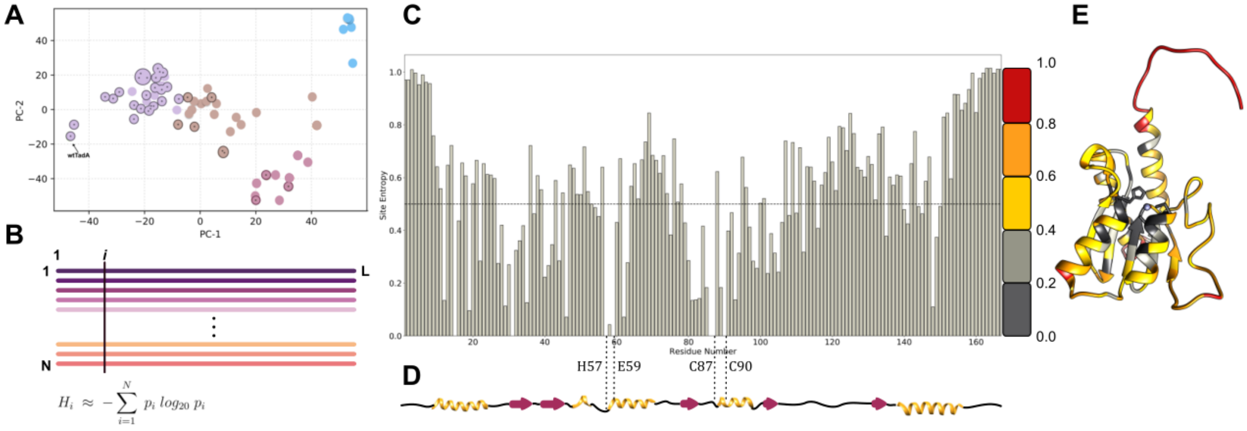


Figure 3.2: **Development and application of sequence-based entropy classifier for TadA.** (A) The top two principal components of the pairwise sequence distance matrix of extant homologs comprising the filtered (indicated with circles and dots) and unfiltered datasets. Based on the similarity of the sequences, the dataset is clustered into 4 separate sets, colored purple, brown, red, and blue. The sequences in the filtered dataset are highlighted in each cluster. (B) Multiple sequence alignment of extant homologs of wtTadA to calculate the statistical probability of occurrence of individual amino acids at residue site i (p_i). This is subsequently used to assign a conservation score to site i, using Shannon's definition of information entropy (H_i) Equation 3.2. (C) Information entropy of individual residue sites of the wtTadA query, with its secondary structure elements mapped below in (D). (E) Entropy values mapped on to the three-dimensional structure of *E. coli* TadA using a color gradient to signify conserved residues and mutational hotspots.

Glu^{59} residue. Consistent with the importance of these residues for the canonical RNA-editing activity of TadA, we observed $\mathrm{H}_i = 0$ for these four active site residues. This active site is further stabilized by a β-sheet core, and the entropy values for 24 of the 38 core residues are also low ($H_i \in \{0.0, 0.4\}$). The surface-exposed residues have relatively higher values of H_i, with the C- and N- terminal residues having the highest values ($H_i > 0.4$) (**Figure 3.2(D)**). By mapping these entropy scores onto the structure of wtTadA[17] (**Figure 3.2(E)**), the correlation between the entropy values and the three-dimensional structure of TadA is clearly apparent. Thus this sequence-based entropy model is capable of representing the structural information encoded by the amino acid sequence of wtTadA.

3.3.2 Sequence Entropy as a Binary Classifier of TadA* Function

Building upon these results, we used the sequence-based entropy model to rationalize the role played by different amino-acid mutations that have been experimentally shown to modulate the function (i.e. RNA-editing activity) of ABEs (**Figure 3.1(B) and (C)**). Based on the biochemical interpretation of the two extreme entropy values, we chose $H_i = 0.5$ as an initial cutoff value to distinguish the functional implications of the entropy data obtained for the wtTadA sequence (**??**C) in the context of all mutations reported for the ABEs (**Figure 3.1(B) and (C)**).[79] Within this model, we hypothesize that residue sites having $H_i > 0.5$ will either induce no change in the activity of wtTadA or, if mutated appropriately, can have a favorable impact on the native activity (i.e., RNA-editing activity) of wtTadA. Conversely, sites with $H_i \leq 0.5$ are predicted to have adverse effects on the canonical RNA-editing activity of wtTadA.[31,41,66] It should be noted that, since our dataset comprises only RNA-editing enzymes, we are primarily referring to the impact that individual mutations have on the native function of the wtTadA sequence (SI sequences and Figure S1B). However, given the vast amount of experimental data available regarding the mutations that impact the ssDNA-editing efficiency of TadA*, we are interested in assessing the performance of the entropy-based model on these mutations as well. Hence, mutations that either increase the ssDNA-editing ability or have no negative impact on the RNA-editing activity of ABEs (as discovered using either directed evolution[11,64,65] or site-directed mutagenesis[41]) are deemed to be correctly classified using our information entropy-based model if their H_i value is greater than 0.5 (**Figure 3.3**). In certain cases residues have been or can be mutated in multiple different ways, which may lead to conflicting editing outcomes (Table S3). To resolve these conflicts and classify such sites, precedence was given to the RNA-editing outcome produced by the most chemically conserved mutation at such sites.

To quantify the performance of the sequence-based entropy model, we computed the confusion matrix where each prediction (**Figure 3.3(A) and (B)**) is validated against the corresponding experimental editing outcome for 45 total mutations[11,31,41,64–66] (**Figure 3.3(C)**). By construction, the diagonal elements of the confusion matrix thus correspond to correct predictions, while the off-diagonal elements indicate incorrect predictions. Despite being entirely derived from the information content of amino-acid sequences contained in a highly biased RNA-editing dataset, the sequence-based entropy model applied to all the reported ABE mutations exhibits a remarkable accuracy of 91.1% and an F_1 score of 0.91 (**Figure 3.3(C)**). Specifically, the model correctly predicted all the mutations that are reported to adversely impact the native RNA-editing activity of TadA*. However, it incorrectly predicted the effects of 4 mutations, all of which correspond to residues with low entropy values that were experimentally found to increase ssDNA-editing activity. Hence, in an attempt to understand the significance of these misclassified residues and identify possible deficiencies of our model so as to refine our classification scheme, we sought to further analyze the amino acid distribution at these residue sites. (**Figure 3.3(A)**). We found that site 82 (Val in the wild-type enzyme) had ambiguous experimental data as its mutation to Gly abrogates RNA-editing in SECURE-ABEs,[41] while its mutation to Ser results in enhanced ssDNA-editing in ABE8s.[64] This suggests higher predictability of the entropy classifier regarding the native RNA-editing activity of TadA* than its ssDNA-editing activity, as expected. Additionally, both sites 84 (Leu in wtTadA) and 108 (Asp in wtTadA) are associated with low entropy values, but were mutated to enhance ssDNA-editing activity during the development of the foundational ABE7.10.[11] Similarly, a low entropy value is found for site 149 (phenylalanine in wtTadA), which was mutated to enhance DNA editing activity in ABE8e.[65] Overall, these misclassifications are restricted to mutations that impact the ssDNA-editing efficiency of TadA*, thus highlighting that the entropy-model, just like other sequence-based coevolutionary methods, is limited by quality of the sequence dataset.[80]

The D108N mutation was the critical first mutation that led to the onset of ssDNA-editing activity of TadA*.[11] Moreover, this residue is part of a surface-exposed loop in the structure of TadA*. Hence, we would expect this residue to display high entropy. To further dissect the anomalous misclassification ($H_i < 0.5$) of site 108 through our entropy-based model, we analyzed the distribution of various amino acids at this site within our dataset (**Figure 3.3(D)**). We observed that, although the mutational entropy of this site is marginally low, approximately 36% of the dataset sequences record an Asn at this site, making it the second most probable amino acid at site 108. This observation was particularly striking given the importance of the D108N mutation; it was the first mutation observed in the directed evolution of the foundational ABE7.10,[11] and we recently discovered that reversion of this mutation in the ABE7.10 construct resulted in complete loss of ssDNA-editing activity by TadA*.[81] It is therefore quite significant that a mutation that is so critical for imparting novel ssDNA-editing functionality onto an RNA-editing enzyme has such a high incidence in naturally-occurring TadA homologs (**Figure 3.3(D)**). Additionally, this indicates that in the case of TadA* evolution, the enzyme achieves activity towards DNA substrates while still retaining activity toward its native RNA targets.

Upon conducting a similar distribution analysis for site 84, which is also misclassified by the entropy-based model as a low entropy site that favorably affects ssDNA-editing, we found that while this core residue has a low sequence entropy of $H_i = 0.42$ as defined by Shannon's entropy, 88.6% of sequences had an aliphatic amino acid (Leu, Val, or Ile) at this position, in direct contrast to its mutation to Phe as in ABE7.10 (**Figure 3.3(D)**). Thus, unlike the D108N mutation, the L84F mutation is a novel mutation that had not been explored by natural protein evolution.

This analysis of the distribution of the possible amino acids based on their chemical nature helps identify the types of mutations that are tolerated at various sites of the TadA* sequence (Figure S2A). Hence, we re-calculated the entropy values for wtTadA by binning amino acid residues according to their side chain classifications: polar uncharged, positively charged, negatively charged, hydrophobic-aliphatic, hydrophobic-aromatic, and special (Gly, Pro). The resulting binned entropy values (Figure S2B, C, and D) were greater than 0.5 for site 108 while still remaining lower than 0.5 for site 84 (**Figure 3.3(D)**). These results thus indicate that the entropy-based analysis allows not only for the quantification of the mutational propensity of individual wtTadA sites but also the characterization of the chemical properties that make mutations to a specific class of amino acids relatively more favorable. Moreover, we also speculate that residue sites having marginally low H_i values can in fact be mutated based on the amino acid distribution observed in its extant homologs to confer novel functionality to the enzyme (as seen for D108N mutation) or to disrupt native functionality (as seen for L84F).

3.3.3 Experimental Analyses

We next sought to experimentally test our hypothesis that the conservation scores and amino acid distributions derived from the entropy-based model could be used to predict the effects of mutations on the RNA-editing activity of TadA*. It is well known that later-generation ABEs induce transcriptome-wide RNA editing, but it is unknown if this is a "carryover" activity from wtTadA being able to edit RNA sequences other than its native tRNA substrate, or if the various mutations identified through directed evolution not only enhanced the ssDNA-editing activity of TadA*, but also its nonspecific RNA-editing activity. We first tested the ability of ABE0.1 (both as monomeric and dimeric wtTadA fused to Cas9n) to introduce A-to-I edits in mRNA in a gRNA-independent manner. We transfected HEK293T cells with constructs encoding monomeric ABE0.1, dimeric ABE0.1, or heterodimeric ABE7.10 (wtTadA-TadA*-Cas9n), extracted mRNA after 36 hours, and used high throughput sequencing to quantify A-to-I editing at six different sites throughout the transcriptome that had previously been shown to be edited by ABE7.10 in a gRNA-independent manner.[41] We observed >50% A-to-I RNA-editing efficiencies at all six sites by both wild-type constructs. Moreover, consistent with the recent report comparing the kinetics of ABEs on RNA substrates,[67] the RNA-editing activity of dimeric ABE0.1 was on average 21% higher than ABE7.10, highlighting the remarkable shift in substrate-preference of wtTadA enzyme due to the many mutations that were found through directed evolution for ABE7.10.

Our entropy-based analysis suggests that non-aliphatic mutations at site 84 would diminish the RNA-editing activity of wtTadA, while certain mutations at site 108 would retain (or even en-

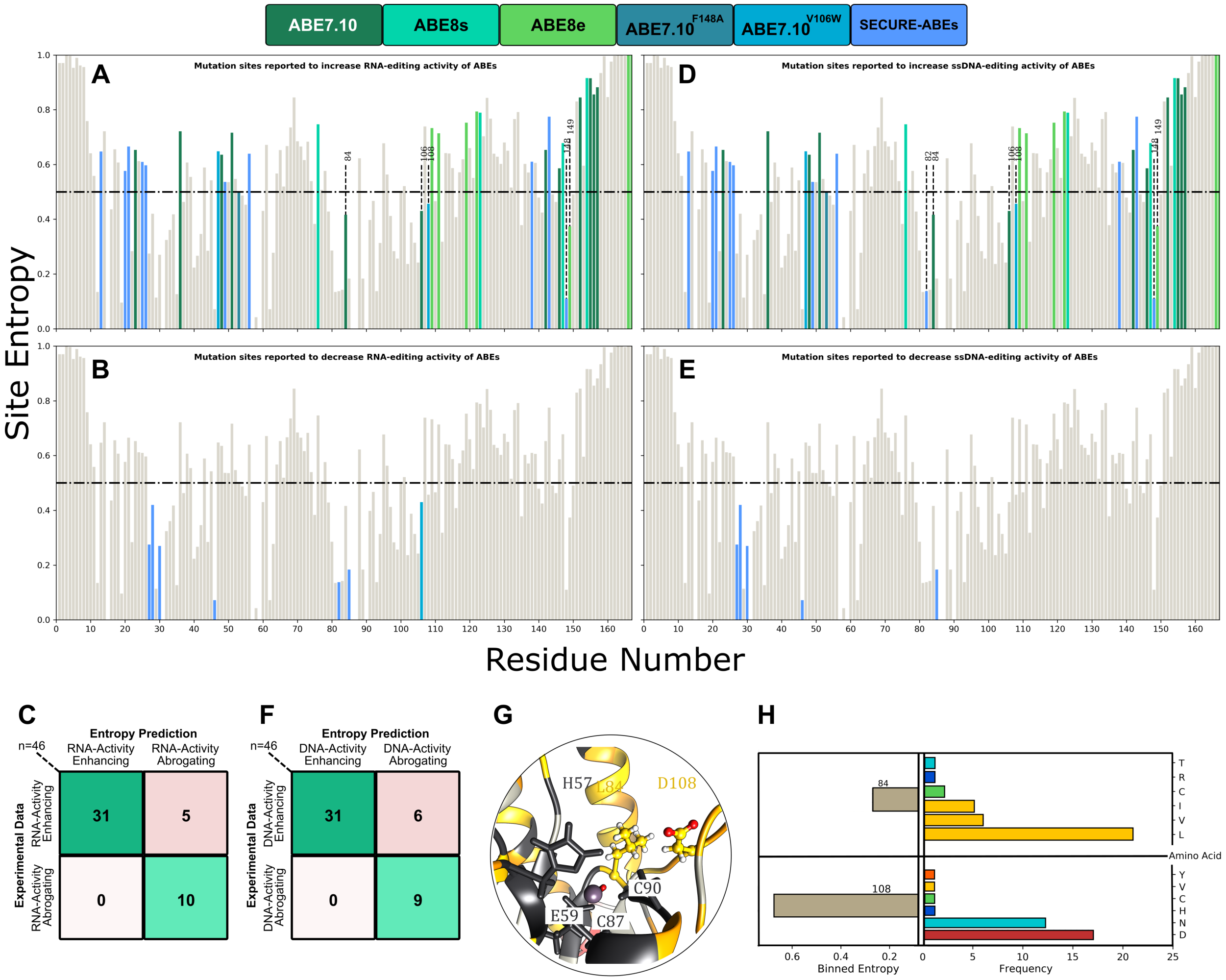


Figure 3.3: **Application of sequence-based entropy classifier to determine the impact of TadA mutations on RNA-/DNA-editing of ABEs.** (A) Mutations reported to have beneficial or neutral effects on the RNA or DNA editing activity of the ABEs. (B) Mutations reported to have detrimental effects on the RNA or DNA editing activity of the ABEs. (C) Confusion matrix of the experimental data and the entropy-based classifier. (D) Binned entropy values and distribution of amino acids at sites 84 and 108. (E) Local environment of 84 and 108 residues in the wtTadA structure.

hance) the RNA-editing activity of wtTadA (**Figure 3.3(H)** and Figure S2). To test this hypothesis, we generated six different ABE variants - ABE1.1 (i.e. ABE0.1(D108N)), ABE0.1(L84F), and ABE1.1(L84F), and their corresponding heterodimeric constructs (i.e. wtTadA-TadA*-Cas9n), and compared their RNA-editing activities with ABE0.1 and ABE7.10 at the same six sites. Each variant was tested as a monomer and a heterodimer to account for any changes in the dimerization ability of TadA due to each mutation.

Consistent with our hypothesis and the entropy-based classification model, the D108N mutation leads to a modest 11.2% (ranging from 5.7% to 21.8%) increase in the A-to-I RNA-editing activity of ABE1.1 compared to ABE0.1. Moreover, the L84F mutation leads to a 25% (ranging from 19.8% to 31.7%), or almost 1.7-fold decrease, in the RNA-editing efficiency of the enzyme as compared to ABE0.1 across the six different RNA sites that were analyzed (**Figure 3.4**). These editing patterns were also observed at an additional UACG motif within RNA site 1, although the editing levels here were much lower than the other six sites (Supplementary Note 2).

This loss of function due to the L84F mutation can be restored either by dimerizing the protein with wtTadA (as ABE0.1(L84F, dimer)) or by adding the D108N mutation (as ABE1.1(L84F)). We speculate that in the case of ABE0.1(L84F, dimer) the observed increase in RNA-editing is due to the addition of the wtTadA subunit, which is capable of efficient RNA editing on its own (as in the ABE0.1 monomeric construct). In the case of ABE1.1(L84F), whose activity is comparable to ABE1.1, we observed a modest 4.3% (ranging from 1.6% to 13%) increase over ABE0.1. This restoration of the RNA-editing efficiency upon the combination of D108N and L84F mutations is particularly interesting as it highlights the non-additive and epistatic effect that mutations can have on enzyme function. Thus, upon combining a high entropy (or high activity) mutation with a low entropy (or low activity) mutation, the resultant double-mutant exhibits high activity, rather than an average of the two activities. Furthermore, this double-mutant exhibits increased activity towards a different substrate (ssDNA).

Intriguingly, the RNA-editing activity of ABE7.10, which has 12 other mutations in

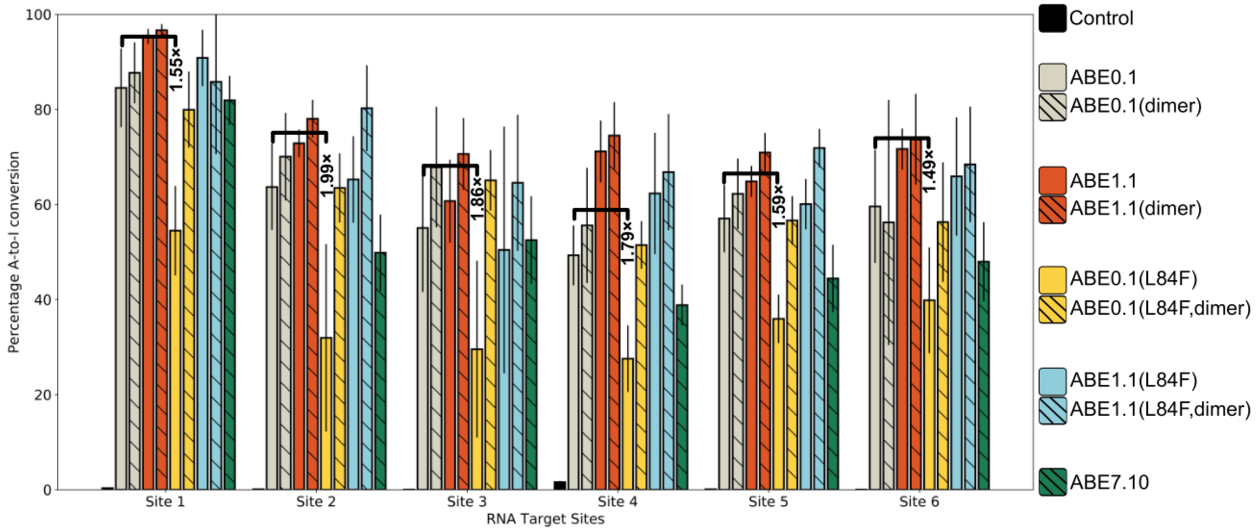
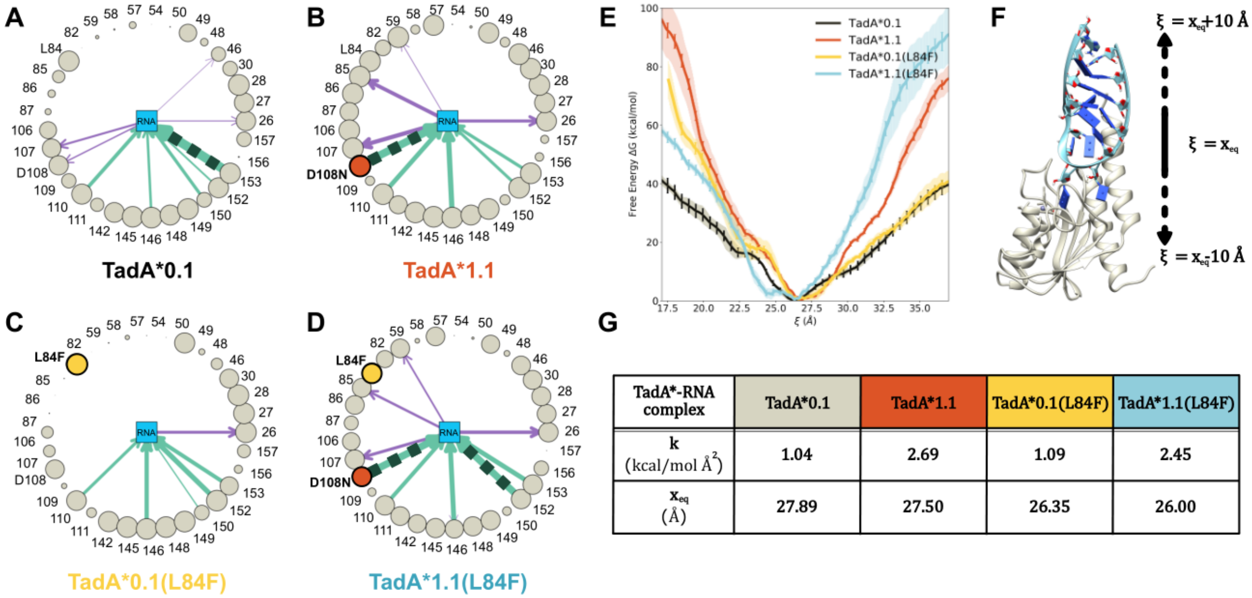


Figure 3.4: **Experimental validation of predictions from sequence-based entropy classifier.** A-to-I base editing efficiencies in HEK293T cells by various ABE mutants at six different gRNA-independent RNA off-target sites. Fold-decrease values associated with the reduction in the RNA-editing upon incorporation of the L84F mutation in ABE0.1 are indicated. Values and error bars reflect the mean and SD of three independent biological replicates performed on different days.

addition to D108N and L84F, is slightly lower than that of ABE0.1 by 8.9% (ranging from 2.5% to 13.8%), or 1.2-fold (**Figure 3.4**). This observation further reinforces the non-additivity of the TadA* mutations identified using directed evolution of ABEs. The early mutations led to a broadening of the substrate specificity of TadA* (i.e., imparting ssDNA-editing capabilities on to TadA) and later mutations enhanced the ssDNA-editing activity while potentially suppressing the RNA-editing activity (as with the L84F mutation).

3.3.4 Interaction and binding of RNA with TadA*

In Ref.,[81] we demonstrated that the effects of individual mutations on the ssDNA-editing activity of ABEs can be studied using a minimalistic model of the system, comprised of the TadA* mutants and the nucleic acid substrates, while ignoring Cas9, which acts as a mere carrier of the nucleotide editing module to its target genomic locus. Moreover, it has been experimentally



TadA*-RNA complex	TadA*0.1	TadA*1.1	TadA*0.1(L84F)	TadA*1.1(L84F)
k (kcal/mol Å^2)	1.04	2.69	1.09	2.45
x_{eq} (Å)	27.89	27.50	26.35	26.00

Figure 3.5: **Analyses of TadA*-RNA contacts and binding.** Asteroid plots for the analysis of the interaction of (A) TadA*0.1, (B) TadA*1.1, TadA*0.1(L84F), and (D) TadA*1.1(L84F) with substrate RNA. (E) Binding affinity comparisons for the various TadA*–RNA complexes. (F) The collective variable (ξ) used to monitor the binding/unbinding of the TadA*–RNA complexes. (G) Parameters associated to harmonic functions fitted to binding energy curves shown in (E).

proven that the off-target RNA-editing by ABEs occurs in a Cas9 (or gRNA)-independent manner, which reinforces the notion that only the TadA* portion of the ABEs act on the RNA off-target substrates[31, 35, 64–66, 82]

To understand the complex epistatic relationship between the L84F and the D108N mutations in the context of the RNA-editing activity of ABEs (**Figure 3.3** and **Figure 3.4**), we modeled the ABE–RNA systems by combining the experimentally resolved structure of wild-type *E. coli* TadA (PDB :1Z3A[17]) and its native 14-mer RNA-hairpin substrate (5′-UUGACU**A**CGAUCAA-3′) (PDB :2B3J[30]). The RNA sequence in our simulation models, as well as the off-target RNA sites that we tested experimentally (**Figure 3.4**), have the same consensus sequence (-UACG-) as that reported previously[35, 41] (Figure S3). Moreover, the structures for these six RNA editing sites resemble the hairpin loop structure of the native target of TadA* that we simulated, further reinforcing the strong preference of TadA* for its native substrate (Figure S4). Having generated these models we then carried out MD simulations for each of the four TadA* mutants–RNA complexes for 1 microsecond and examined the trajectories for changes in interactions between individual TadA* residues and the nucleic acid substrate. Since the mutations we are interested in lie near the active site of the TadA*, we focus predominantly on the interactions between the nucleotide bases splayed in the active site, i.e., the target adenine and its 5′ and 3′ flanking bases (U**A**CG) and neighboring TadA* residues. To hone in on the amino acids in direct contact with these nucleotides, we carved out a 4 Å search radius around these bases, and then project the residues that lie within this sphere onto asteroid plots (**Figure 3.4(A) to (D)**). In the asteroid plots, the nucleotides in the active site are represented collectively as the central node, and the peripheral nodes correspond to all amino acids within the first interaction shell of the nucleotides in the active site. The size of the encircling nodes is proportional to the time that the corresponding residues spend within the first interaction shell of the RNA bases throughout the entire MD trajectory. The hydrogen bonds (H-bonds) between these residues and the RNA bases are depicted as arrows connecting the relevant nodes in each asteroid plot, with the thickness of the arrows being proportional to the stability of the H-bond itself, which is defined as the frequency of appearance of that H-bond during the simulation. The comparisons between the TadA*0.1/TadA*0.1(L84F), TadA*0.1/TadA*1.1, and TadA*1.1/TadA*1.1(L84F) mutants indicate that the D108N mutation leads to the formation of a favorable H-bond between the Asn108 residue and the U base flanking the target **A**. In fact, the TadA*1.1(L84F) mutant has the strongest interaction with RNA as the L84F mutation causes additional structural rearrangements surrounding the active site resulting in a double H-bond interaction with the RNA substrate through residue 152. The weak H-bond between D108 and the 2′-OH group of the flanking U base predicted by our simulations is also found in the crystallographic structure of the wtTadA-tRNA complex (PDB ID: 2B3J[30]). However, this weak H-bond does not appear in the TadA*0.1(L84F) mutant and is replaced by a much stronger H-bond in the TadA*1.1 mutant upon mutation of glutamate to asparagine at site 108. The formation of this stronger H-bond also leads to an increase in interactions between some of the peripheral residues (57, 59, 82, 85, 86, and 87) and the RNA bases in the active site, indicating a more stable conformation of the target adenine. A similar increase in the interaction induced by the H-bond formed by the D108N residue was also observed in our MD simulations of the TadA*–ssDNA complex.[81] In the context of ssDNA-editing by ABEs, the D108N mutation in ABE1.1 leads to the onset of activity on DNA via the formation of this H-bond donation.[11,81] However, in the context of RNA-editing efficiency, the D108N mutation, and consequent formation of the H-bond with RNA, only amounts to a slight increase in the activity due to wtTadA (ABE0.1) being already highly proficient in editing its native RNA substrate, as well as ssRNA in general (**Figure 3.4**).

Although the L84F mutation, unlike D108N mutation, is accompanied by a more pronounced effect on the RNA-editing activity of wtTadA (**Figure 3.4**), and is in fact a novel mutation in the first interaction shell of the RNA bases (**Figure 3.3(D)**), the comparison of the asteroid plots corresponding to TadA*0.1 and TadA*0.1(L84F) shows less drastic changes, than those observed in the TadA*1.1 asteroid plot. Specifically, the L84F mutation leads to the elimination of the weak H-bonds established by the 107 and 108 residues in TadA0.1. To quantify these differences, we performed umbrella sampling (US) simulations to determine local changes in the binding free energies of the various TadA*–RNA complexes in the neighborhood of the active site. Starting from the equilibrated structure of each TadA*–RNA complex, we modeled the binding process using a collective variable (ξ) defined as the distance between the TadA* and RNA centers of mass, which was varied from 17 to 37 Å. We successfully used the same collective variable to characterize the binding process in the analogous TadA*–DNA complexes in Ref.[81] The free-energy changes along this collective variable were calculated for each of the four TadA* mutants using the weighted histogram analysis method (WHAM).[51,83] **Figure 3.5(E)** shows that the TadA*1.1–RNA and TadA*1.1(L84F)–RNA complexes are more tightly bound than the TadA*0.1–RNA complex. While these trends help explain the experimental RNA-editing efficiencies of these D108N mutants when compared to ABE0.1, they do not apply to the TadA*0.1 and TadA*0.1(L84F) mutants which exhibit similar local free-energy changes as

RNA is pulled out of the active site. This observation reciprocates the results of our previous study of the TadA*–ssDNA complex showing that mutations installed at later stages of the directed evolution process do not further enhance the binding strength relative to TadA1.1, but instead most likely impact the catalytic activity of TadA*.[81]

As the subtle conformational changes that we observe between the asteroid plots of TadA*0.1 and TadA*0.1(L84F) (**Figure 3.5(A)** and **(C)**) do not result in significant changes in the binding strength between these two mutants, we thus sought to quantify the effects of these conformational changes on the catalysis.

3.3.5 Water in the active site and implications for catalysis

The hydrolytic deamination reaction catalyzed by TadA involves a zinc-coordinated water molecule (hereafter referred to as the activated water molecule) that is deprotonated by the active site Glu^{59} residue (a highly conserved residue, see **Figure 3.2**) during the first step of the reaction. In addition to this activated water molecule, the active site also includes another structurally important water molecule (hereafter referred to as the bridging water) which acts as a bridge between Glu^{59} and the carbonyl backbone of Leu^{84} (**Figure 3.6 A and B**). Both water molecules are resolved in several high-resolution crystal structures of TadA homologs (Figure S5) which further reinforces their importance in the stabilization of the active site cavity.

To characterize the role played by these two water molecules in the active site of the various TadA*–RNA mutant systems, we analyzed the data from our MD simulations in the form of modified chord diagrams in (**Figure 3.6 C, D, E, and F**). The persistence of the activated water molecule is depicted in red in the left partitions while that of the bridging water is depicted in blue in the right partitions of the four panels. The thickness of each chord is proportional to the time spent by the corresponding water molecules in the active site (Figure S6). For the TadA*0.1–RNA and TadA*1.1–RNA systems, we found that these two water molecules are highly stable in their respective positions and do not undergo any diffusion throughout the entirety of our

MD simulations (1 μs). In contrast, for the TadA*0.1(L84F)–RNA system, both water molecules exhibit higher mobility and are exchanged several times with water molecules initially located in the bulk solution at the beginning of the simulation. We speculate that the hydrophobic-aromatic nature of the phenylalanine residue may be responsible for the decreased stability of both water molecules in the active site (Figure S7). The stability of the two water molecules is restored in the TadA*1.1(L84F)–RNA complex. This implies that the D108N mutation is capable of canceling out the destabilizing effects of the L84F mutation and effectively modulating the hydration of the active site, despite not engaging in any direct contact with either water molecules. We observe similar trends when comparing these mutations in the apo-TadA* simulations. Specifically, the apo-TadA*0.1(L84F) system shows an analogous increased flux of the two water molecules in the active site, which is again suppressed after the installation of the D108N mutation (Figure S8). Importantly, the changes in the persistence of these catalytically-relevant water molecules in the active site of the TadA*–RNA/TadA* systems (**Figure 3.6 C, D, E** and, **F** and S8) mirrors the changes in RNA-editing activity measured for the ABEs (**??**).

Since the first step of the adenine deamination reaction involves the deprotonation of the activated water molecule by the Glu^{59} residue, we speculate that the changes we observe in the stability of the active site water molecules may lead to changes in the reaction rates in the four TadA* mutants. Hence, for a more explicit comparison with the experimental catalytic data of these four TadA* mutants, we performed quantum mechanics/molecular mechanics (QM/MM) simulations to investigate the first step of the reaction mechanism. Owing to the high computational cost of simulating the entire system at the QM level, QM/MM simulations offer an ideal trade-off between accuracy and computational efficiency by simulating the reaction centers with QM accuracy while the remaining system is treated at the MM level of theory. In our QM/MM simulations, the QM region encompasses the side chains of active site residues (His^{57}, Glu^{59}, Cys^{87}, Cys^{90}), Zn^{+2}, and the activated water molecule and are treated at the DFTB3 level with 3OB parameterization,[84–86] while all other atoms of the system are included in the MM

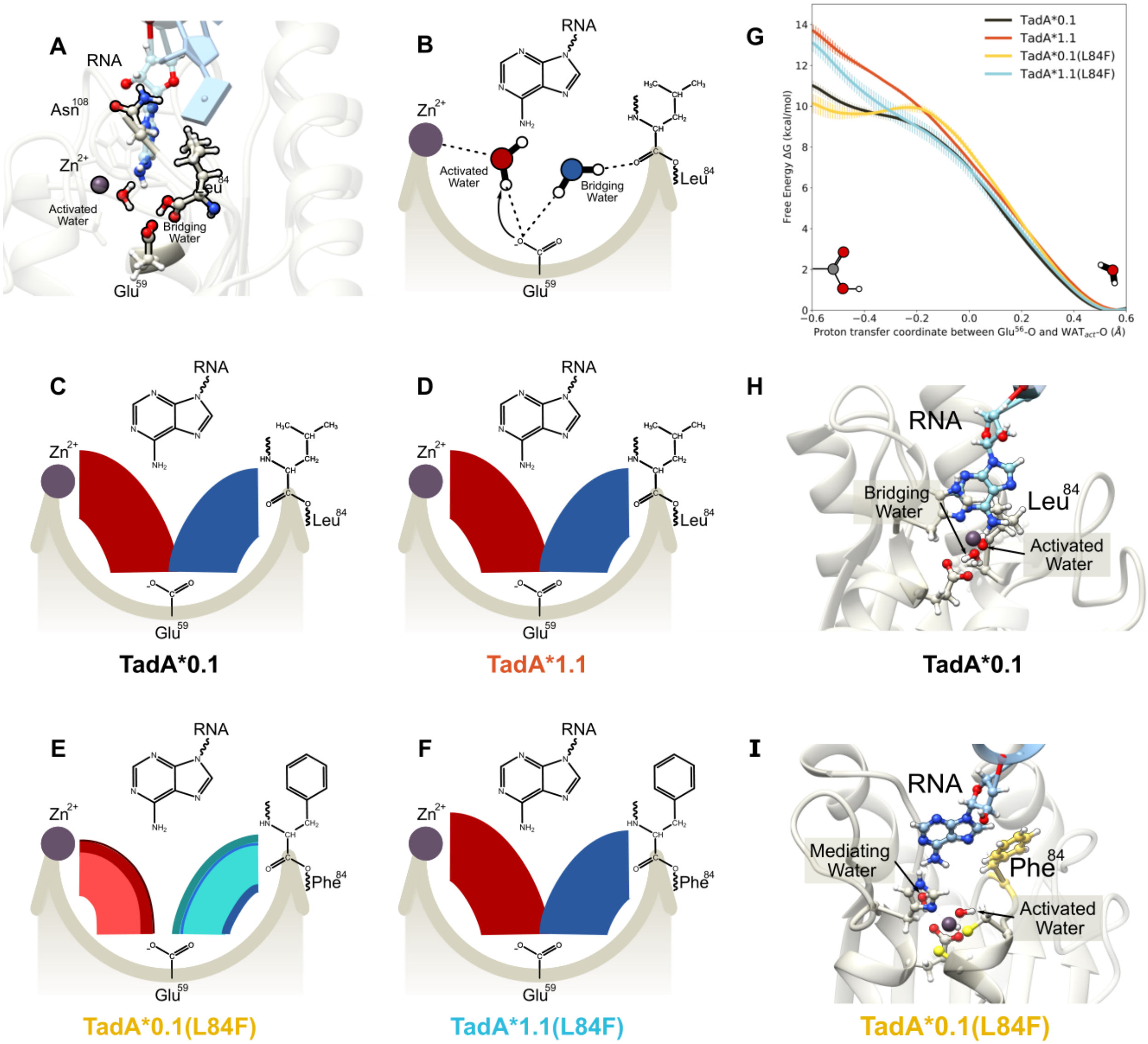


Figure 3.6: **Analyses of active site waters their role in the RNA-editing catalysis by TadA*.** (A) Side view of the of the TadA*–RNA system highlighting the location of the catalytically relevant residues. The Zn^{+2} ion is coordinated by His^{57},Cys^{87}, and Cys^{90} (not shown here for clarity) and a water molecule. This water molecule is activated by Glu^{59}, which is also connected to another water molecule. This second water acts as a bridge between the Glu^{59} and the carbonyl backbone of residue 84. The target adenine is deep within the active site and residue 108 is farther away from the active site waters. (B) Simplified flat lay representation to highlight the interactions of active site waters. Modified chord diagrams to demonstrate the persistence of the active site waters for (C) TadA*0.1–RNA, (D) TadA*1.1–RNA, (E) TadA*0.1(L84F)–RNA, and (F) TadA*1.1(L84F)–RNA. (G) Reaction profile for the deprotonation of the activated water molecule the various TadA*–RNA systems. (H) Conformation of the TadA*0.1–RNA when the proton resides on Glu^{59}. (I) Conformation of the TadA*0.1(L84F)–RNA when the proton resides on stability on the Glu^{59}. The target **A** has moved back into the active site, towards the Phe^{84} and is separated from the active site residues by an additional water molecule - the mediating water.

region. Similar DFTB approaches have been successfully employed in the past to study several zinc-containing enzymes,[87–90] including deaminases which are homologous to TadA.[91–93]

All QM/MM simulations were initiated from configurations taken from the US windows corresponding to the PMF minima shown in **Figure 3.5**. In modeling the first step of the deamination reaction, configurations with undissociated activated water molecules define the reactant state, while configurations with the protonated Glu^{59} residue define the product state. In the transition state, the proton is equally shared by the activated water molecule and Glu^{59}. To determine the energetics associated with this proton-transfer reaction, QM/MM umbrella sampling simulations were carried out along the proton-transfer coordinate, which is defined as the difference between the distances of the shared proton from the activated water and Glu^{59} as used in Ref.[94] The PMFs calculated using WHAM for all four TadA* mutants are shown in **Figure 3.6(G)** and are consistent with the energetics observed for other zinc-containing deaminases calculated using various QM/MM methods (*E. coli* CDA,[95] yeast CDA,[94,96] guanine deaminase[97]).

The PMF profiles indicate that only the TadA*0.1(L84F)–RNA complex is associated with a weakly stable product state (i.e. a protonated Glu^{59}). At first glance, these results seem to be counter-intuitive and contrary to the experimental observation of a lower RNA-editing activity for the ABE0.1(L84F) mutant. However, upon further examination of the product state in the TadA*0.1(L84F)–RNA complex, we observed that proton transfer from the activated water to Glu^{59} is accompanied by the concomitant movement of the target adenine base away from the active site residue and towards the aromatic phenylalanine ring deeper into the active site forming in a staggered pi-stack with L84F residue (Table S1). The cavity formed as a result of this conformational change of the target adenine is filled by another water molecule (hereafter referred to as the mediating water) that may contribute to the following reaction steps, thereby altering the reaction mechanism for TadA*0.1(L84F). In this context, it should be noted that a deamination mechanism involving extra water molecules was characterized for cytosine deaminase in Ref.[98]

In the case of the TadA*0.1 (or TadA*1.1 and TadA*1.1(L84F)) system, our simulations predict that the active site retains its configuration, with the adenine base primed for the next steps of the reaction (**Figure 3.6(H)**). We thus hypothesize that the proximity of the adenine base prevents the transfer of the proton from the activated water molecule to Glu^{59}. QM/MM umbrella sampling simulations carried out for the apo-TadA* mutants provide support for this hypothesis, showing the formation of a stable product state for all the TadA* mutants due to the lack of the adenine base (Figure S8G).

We thus conclude that the L84F mutation, a novel mutation at a low entropy residue site, affects the decrease in the activity of TadA* on the native RNA substrate through two key changes in its deamination chemistry. First, this mutation destabilizes the two water molecules in the active site, which are both structurally and functionally relevant to the initiation of the deamination reaction. Second, it pulls the target adenine base away from the protonated Glu^{59}, thereby making the subsequent reaction steps less feasible or leading to an alternate reaction pathway involving additional steps (e.g., through the mediating water). Our simulations indicate that the combination of the D108N mutation, which increases the RNA binding affinity, with the L84F mutation conserves the integrity of the active site by both stabilizing the two water molecules and positioning the target adenine appropriately for subsequent reaction steps, thereby, rescuing the catalytic activity of the ABE1.1(L84F) mutant (**Figure 3.4**).

3.4 Discussion

Through a systematic investigation of the various mutations that have been thus far identified in TadA*, our study re-traces the evolutionary trajectory followed by this enzyme using a data-driven approach that combined statistical models, MD simulations, and experimental assays.

The information contained in the naturally-evolved TadA homologs aids in rationalizing the effects of the mutations that have accumulated in the laboratory-engineered TadA* (**Figure 3.2**). We have demonstrated that mutations with a favorable impact on the RNA-editing activity of TadA* occur at residue sites having higher entropy, whereas mutations with an unfavorable impact on the RNA-editing activity occur at residue sites with lower entropy (**Figure 3.3**). Moreover, these low entropy sites when mutated to previously unvisited amino acids in the sequence space, such as the L84F mutation, can also have an adverse impact on the native function of the enzyme. Our experimental analyses also revealed that ABE0.1 has remarkably high gRNA-independent off-target RNA-editing and is even higher than the evolved ABE7.10 variant (**Figure 3.4**).[41,64,65,67] These results indicate that such entropy-based scores, albeit being extracted from a highly RNA-biased dataset, can serve as a preliminary screen for site-directed mutagenesis and guide the library preparation for evolving future base editors with reduced off-target transcriptome editing activity. The most reliable inferences that can be derived from such biased datasets are related to the native RNA-editing functionality of the query sequence. Hence, we propose that this entropy-based tool be preferentially applied for the search of mutations that can suppress the inherent RNA-editing activity of potential base editors, a problem that, at present, cannot be solved using the traditional directed evolution methods.

3.5 Methods

3.5.1 Data Curation and Sequence Entropy

Extant homologs were obtained using BLAST program[74] using E. coli wtTadA as the initial query sequence with an e-value cutoff of 0.1 in the SWISSPROT database.[71] We further filtered the dataset, by removing sequences with more than 40% gap percentage and to minimize redundant sequences with more than 95% identity to the query sequence. The final filtered dataset comprises of 35 homologs. The resultant dataset was used to calculate the sequence entropy score, defined as follows:

$$H_i \approx -\sum_{n=1}^{N} p(i_n)\, log_{20}\, p(i_n) \quad \text{for} \quad i \in \{1,\ldots,L\} \tag{3.2}$$

where p(i_n) refers to the statistical probability of having a particular amino acid *n* at site *i* and N is the total number of amino acids. Further details regarding the dataset and entropy calculation can be found in the Supplementary Materials and Methods.

3.5.2 Computer simulations

The TadA*0.1 model was built using the crystal structure of E. coli TadA (PDB ID: 1Z3A).[17] Given the sequence homology between S. aureus TadA and E. coli TadA, we combined the saTadA-RNA structure (PDB ID: 2B3J) with the TadA*0.1 model to build the TadA*0.1-RNA model26. The TadA*0.1 was transformed into the various ABE mutants using the swapaa command in Chimera.[42] For both apo-TadA* and TadA*-RNA models, all crystallographic water molecules within 3 Å distance of the surface of the protein or the RNA were preserved during the modeling procedure. All titratable residues were protonated using the H++ server employing the default settings.[44,99]

The protein was represented using Amber ff14SB[45] and the RNA was represented using RNA.OL3 force field.[100–102] The metal-containing active site of TadA* was represented with custom force field parameters obtained using the MCPB.py approach at B3LYP/6-31G* level of theory.[103] LEap tool from AmberTools was used to immerse the apo-TadA* and TadA*-RNA complexes into a pre-equilibrated truncated octahedron box of explicit TIP3P water, with a 15 Å buffer distance. Varying number of Na+ ions were added to each of the systems to maintain electroneutrality and the simulation cell was then replicated infinitely in three dimensions to impose periodic boundary conditions. All MD simulations were performed under periodic boundary conditions using the CUDA accelerated version of PMEMD implemented in Amber18 suite of programs.[47–50] The structures were first relaxed using a combination of steepest descent and conjugate gradient minimization. This was followed by a 1ns heating to 298.15 K and multi-step equilibration under progressively decreasing harmonic restraints for 40 ns. Subsequently, we removed all restraints and carried out 1μs unbiased MD simulations for the four TadA* mutants and corresponding TadA*-RNA complexes.

We calculated the free-energy binding profiles of the TadA*-RNA complexes along the collective variable corresponding to the distance between the centers of mass of the protein and the RNA substrate. For each TadA*-RNA complex, the PMF along this collective variable was calculated using umbrella sampling simulations.[104] Starting from the equilibrated TadA*-RNA structures, we conducted four independent sets of umbrella sampling simulations for all the four TadA*-RNA complexes, and the final PMFs were reconstructed using the weighted histogram analysis method (WHAM) algorithm.[51,83] Additional error analysis was carried out using a custom block averaging script based on the method described by Zhu and Hummer.[52]

The free energy changes for the deprotonation of the activated water molecule by the Glu59 residue for the TadA* (and TadA*-RNA) models were computed for the various systems through a hybrid quantum mechanical/molecular mechanical (QM/MM) approach. The QM subsystem consisted of the side chains of the active site residues (His57, Glu59, Cys87, and Cys90), the Zn^{+2} ion, and the activated water for both the apo-TadA* and TadA*-RNA models. These QM atoms were treated using self-consistent charge density functional tight binding (SCC-DFTB) method implemented within Amber18.[87] The atoms beyond this active site cluster were represented the MM subsystem and were treated using the force fields as in the unbiased MD. The difference of the distances between the active site water oxygen atom and shared proton and the Glu59 O and shared proton, was chosen as the collective variable to monitor the deprotonation reaction. For both the apo TadA* and TadA*-RNA complexes the reaction profile along this collective variable was calculated using umbrella sampling simulations following a procedure similar to the one employed for the calculation of the TadA*-RNA binding profiles as summarized above.

The CPPTRAJ module implemented within Amber18 was used to analyze all the MD trajectories.[28,29] The visualization of the MD trajectories was rendered using Chimera, and data were plotted using Matplotlib.[54] Additional details for all simulation protocols as well as experimental protocols can be found in Supplementary Materials and Methods. Table S3 summarizes all the simulations that were carried out during this study.

3.6 Acknowledgments

The authors thank M. Norman for a Director's Discretionary Allocation on the Comet GPU cluster at the San Diego Supercomputer Center. K.L.R. thanks C. Egan, E. Lambros, and S. Gu for helpful discussions. This research was supported by the University of California San Diego and the NIH through grants no. 1R21GM135736-01 and 1R35GM138317-01. All computer simulations used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF through grant no. ACI-1548562. B.L.R. is supported by the Chemistry-Biology Interface (CBI) Training Program (NIGMS, 5T32GM112584).

Chapter 3 is reproduced, in part, with permission, from: Rallapalli, K.L., Ranzau, B.L., Ganapathy, K.R., Paesani, F., Komor, A.C. 2022. Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors. *CRISPR J.* 5:294–310. The dissertation author was the primary author on all reprinted materials.

Chapter 4

Understanding the effects of remote mutations in ABEs

4.1 Abstract

During the course of evolution of Adenine Base Editors(ABEs), most beneficial mutations that were discovered in its DNA-editing module, TadA, lie far away from the target adenine base itself. The role of these distal mutations in enhancing the editing activity of ABE remains unclear and contrasts most traditional structure-guided methods for protein engineering. In this study, we build upon the recently resolved structure of the ABE and advancements in computational protein structure prediction softwares, to model the conformational dynamics of the TadA domain in the context of the entire Cas9 R-loop complex. Through multi-microsecond simulations of these large biomolecular complexes we elucidate the contributions of the distal mutations of TadA in expanding the substrate scope of the enzyme. We show that despite lying farther away from the active site, these mutations influence the interaction between the nucleotides flanking the target base and the deaminase domain, thereby expanding the substrate specificity of the ABEs.

4.2 Introduction

The goal of efficient A•T→G•C base pair conversion at precise genomic targets has now been achieved through the advent of the Adenine Base Editor(ABE) technology. ABEs consist of an impaired Cas9(Cas9n) enzyme, which acts as the DNA-targeting module, covalently fused to an evolved variant of tRNA adenosine deaminase enzyme, TadA*, which acts as the DNA-editing module. While wild-type TadA(wtTadA) natively deaminates adenines in the hairpin motif of tRNA anticodon stemloop,[17,55] it shows no activity on the single-stranded portion of the DNA exposed by Cas9n during the R-loop formation[11](**Figure 4.1A,B**). Extensive protein engineering and directed evolution led to the discovery of 14 mutations in the TadA* domain, henceforth referred to as TadA*7.10, which enhanced the ssDNA deamination efficiency of the foundational ABE, the ABE7.10.[11] More recently, further evolution and engineering of the ABE7.10, has led to the discovery of six additional mutations (that is 22 mutations in total) in the TadA* domain, henceforth referred to as TadA*8e, which enhances the ssDNA editing activity of the ABE8e variant even more[65](**Figure 4.1C, D**). Intriguingly, many of these activity enhancing mutations lie much further away from the target Adenine itself(**Figure 4.1E, F**). In fact, 19 of the 22 mutations in TadA*8e lie greater than 12 Å away from the target adenine base. At first, this may seem surprising, especially from a structure-base protein engineering perspective. However, given that majority of the amino acids in any protein are on its surface rather than the active site and as directed evolution as it entails random mutagenesis of the entire enzyme, this observation is very common for many evolved enzymes.[105]

In case of ABEs evolution, a unique pattern emerges. Almost 50% of all ABE8e(and 50% of all ABE7.10) mutations occur in the C-terminal α-helical region, α5 helix, of TadA*. This terminal region also links TadA* to Cas9n via a long and flexible XTEN linker(**Figure 4.1D**). How these distal mutations in the terminal helical region of TadA* are enhancing the activity of the enzyme remains unclear, even though we now have a cryo-EM derived structure of ABE8e in action.[67] To further our molecular understanding of these remote α5 helix in the context of the entire ABE8e structure, we have carried out a large scale computational modeling and simulations of ABE0.1, the wtTadA variant of ABE, and compared it with the dynamics for ABE8e.

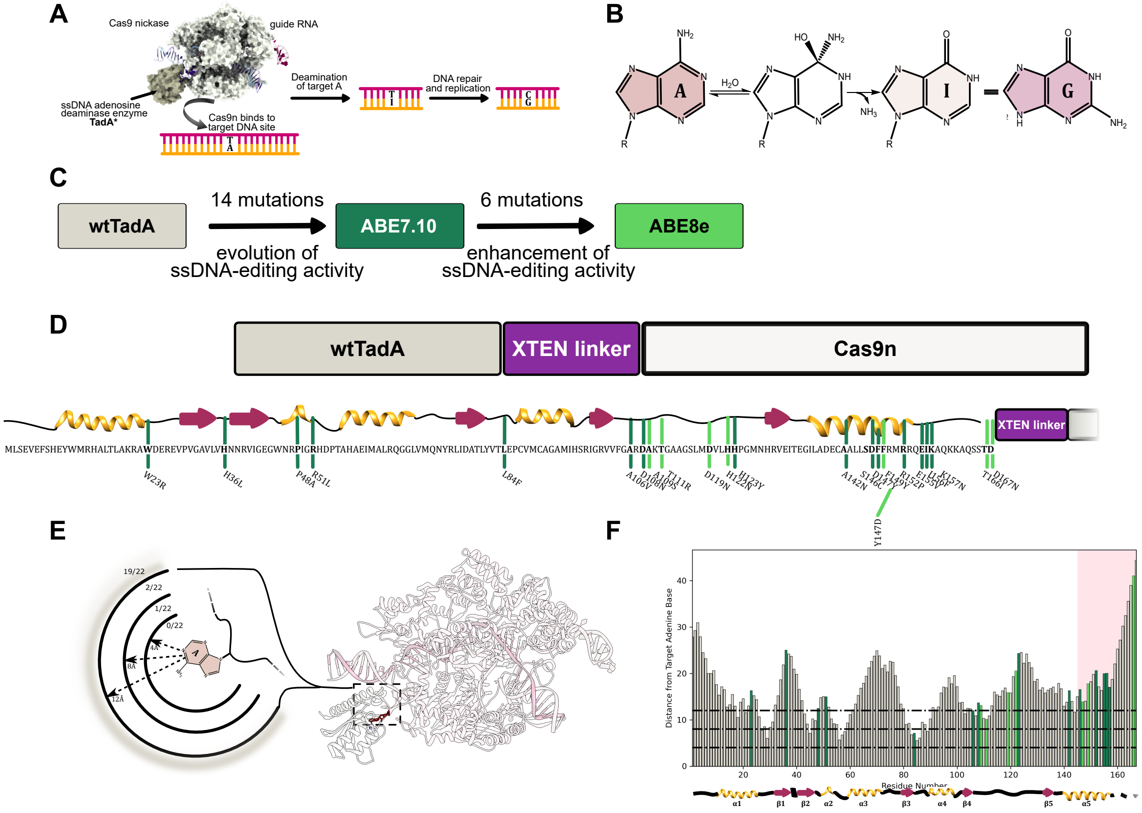

A
Cas9 nickase
guide RNA
ssDNA adenosine deaminase enzyme TadA*
Cas9n binds to target DNA site
Deamination of target A
DNA repair and replication
B
A
I
G
H_2O
NH_2
NH_3
C
wtTadA
14 mutations
evolution of ssDNA-editing activity
ABE7.10
6 mutations
enhancement of ssDNA-editing activity
ABE8e
D
wtTadA
XTEN linker
Cas9n
MLSEVEFSHEYWMRHALTLAKRAWDEREVPVGAVLVHNNRVIGEGWNRPIGRHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTLEPCVMCAGAMIHSRIGRVVFGARDAKTGAAGSLMDVLHHPGMNHRVEITEGILADECAALLSDFFRMRRQEIKAQKKAQSSTD
W23R
H36L
P48A
R51L
L84F
A106V
D108N
A109S
T111R
D119N
H122N
H123Y
A142N
S146C
D147Y
F149Y
R152P
E155V
I156F
K157N
T166I
D167N
Y147D
XTEN linker
E
19/22
2/22
1/22
0/22
4Å
8Å
12Å
A
F
Distance from Target Adenine Base
Residue Number
0
10
20
30
40
20
40
60
80
100
120
140
160
α1
β1
β2
α2
α3
β3
α4
β4
β5
α5

Figure 4.1: **Remote Mutations in ABEs.** (A) A schematic representation of base editing by ABEs(PDB ID: 6VPC[67]). The ABEs studied as a part of the current work consist of a Cas9n fused to an evolved TadA* protein. The binding of Cas9n to the target genomic locus unwinds the DNA double helix and exposes a small region of ssDNA. TadA* acts on this ssDNA and deaminates adenine (A) to form inosine (I), which is subsequently converted to guanine (G) through DNA repair and replication. (B) Overall chemical reaction catalyzed by ABEs. (C) Evolutionary trajectory of ABEs. (D) Primary and secondary structure of ABE8e, with key mutations in TadA* highlighted. The line colors correspond to colors shown in (C), indicating the ABE version in which these mutations were identified. (E) and (F) Distance of TadA*8e mutations from the target Adenine base, calculated based on the recently resolve structure.[67]

4.3 Results

4.3.1 Differences in flexibility of α5 helix and linker

To decipher the role played by the distal mutations in the α5 helix region of TadA* we initiated our investigations by comparing the conformational dynamics of the ABE0.1 and ABE8e complexes(**Figure 2.2** A and B). We performed 2μs long all-atom MD simulations starting with the cryoEM structure of ABE8e[67] to gain insights into the structural dynamics of the system and compared it to the dynamics of the unmutated ABE0.1 model(see Methods). The simulations indicated that the ABE0.1 has a rigid α5 helix terminating the TadA*0.1 domain that is connected to a highly structured XTEN linker region(**Figure 2.2** C and D). This rigidity in the ABE0.1 structure is contrasted by the ABE8e, as the α5 helix has higher flexibility as does the XTEN linker connecting TadA*8e to Cas9n(**Figure 2.2** C and D). It should be noted that the only differences that are observed in the flexibility of the TadA* domains of these two ABE variants occur in the C-terminal helix region of the protein. Given the high density of amino acid mutations in this terminal region of the protein, we speculate that these mutations are in fact changing the rigid nature of the α5 helix into a an extension of the linker region.

4.3.2 Differences in substrate interaction

Next, we sought to correlate the differences observed in the conformational dynamics of the α5 helix region of TadA*s with the substrate engagement in the active site of the enzyme. TadA*, like other members of the CDA superfamily of enzymes, has a characteristic substrate binding active site. When TadA*, binds to its native RNA substrate, three bases in the anti-codon loop(consensus sequence 5′-T**A**CG-3′) of tRNA hairpin structure get splayed into the active site groove of the enzyme, with the target adenine occupying the central position adjacent to the zinc-coordinated active center.[17] The nucleotide upstream(that is the -1 position with respect to the target adenine, T for TadA* RNA substrate) of the target base interacts with the active

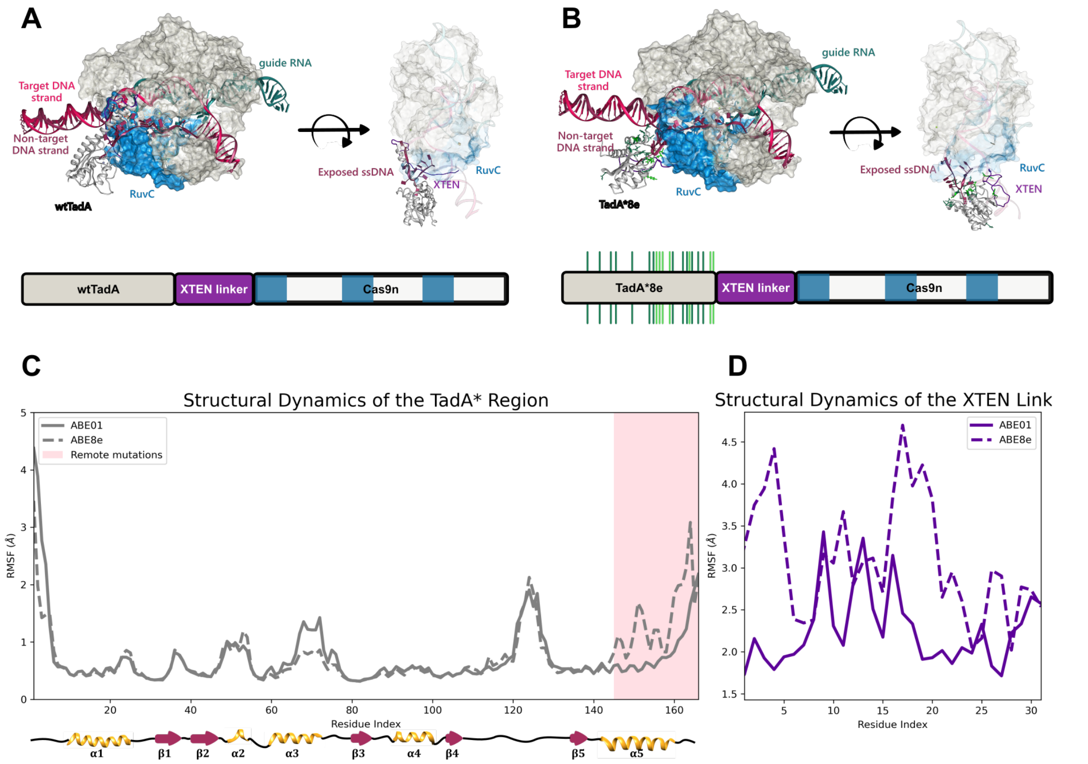


Figure 4.2: **Changes in conformational dynamics of the α5 helix due to remote mutations.** Structural representation of ABE0.1(A) and ABE8e(B) studied as a part of the current work. The mutations in TadA*8e are highlighted as green sticks. TadA*0.1 and TadA*8e bound to the target ssDNA are highlighted to indicate the changes in the conformations of the a5 helix region of TadA*. (C) Residue level flexibility of TadA* shown in terms of the root mean squared fluctuation (RMSF) of the Cα atoms of the peptide backbone. The α5 helix region is highlighted in pink. (D) RMSF fluctuation of Cα atoms of XTEN linker.

site loop connecting the β4-β5 strands and the nucleotide downstream(that is the +1 position with respect to the target adenine, C for TadA* RNA substrate) interacts with the active site loop connecting the α1-β1 strands and is wedged between this loop and the α5 helix.[17]

To gain molecular insights into the interaction between ssDNA (substrate5′-CACT-3′) and TadA*, we projected the interactions between the target adenosine and its flanking bases and the surrounding amino acids onto asteroid diagrams (**Figure 4.3 A and B**). In these diagrams, we use a network representation in which these three nucleotides of the DNA are depicted as the central node and the TadA* residues are the peripheral nodes. As the typical donor atom–donor hydrogen acceptor atom distance is approximated to be 3.5 Å in globular proteins,[26,27] we defined the first interaction shell around the DNA as all amino acids within 4 Å of the three bases in the active site. The size of each node is proportional to the time individual residues spend within the 4 Å shell during the simulation. Hydrogen bonds between residues [defined as in the CPPTRAJ package[28,29]] are depicted as arrows connecting the corresponding nodes, with the arrow size being proportional to the hydrogen-bond strength, which is defined as the number of times that the specific hydrogen bond is established.

As seen in or previous studies of the minimalistic system of TadA* variants bound to ssDNA substrate[81] as well as RNA substrates,[106] we once again observed an increase in the hydrogen bonding between the -1 base and the β4-β5 loop region. Additionally, we observed a hydrogen bond between the α1-β1 active site loop and the +1 base in the ABE0.1 simulations, which is not seen in the ABE8e simulations. This hydrogen bond between the α1-β1 active site loop and the +1 base of the substrate ssDNA is reminiscent of the manner in which TadA*0.1 engages RNA substrates and achieves substrate specificity. Intriguingly, this hydrogen bond is not observed at the start of the ABE0.1 simulation, as the +1 base lies perpendicular to the extended α5 helix due to the manner in which our initial model was generated (see Methods). However, due to the rigidity of the unmutated α5 helix in ABE0.1 compared to the kinked α5 helix in the ABE8e structure(**Figure 4.3D**), the +1 base undergoes a significant conformational drift towards the α1-β1 active site loop, adopting a conformation similar to the native RNA consensus sequence of wtTadA(**Figure 4.3C**).

4.3.3 Differences in interaction with the exposed ssDNA substrate

To determine the cause for differences in the conformational flexibility of ABE0.1 and ABE8e α5 helix we furthered our analysis beyond the three bases in the active site. The activity window of canonical ABEs extends from 4-8 bases downstream of the PAM sequence, GGG, base pair in our simulations (**Figure 4.4 A**). To fully analyze all the interactions that the target exposed

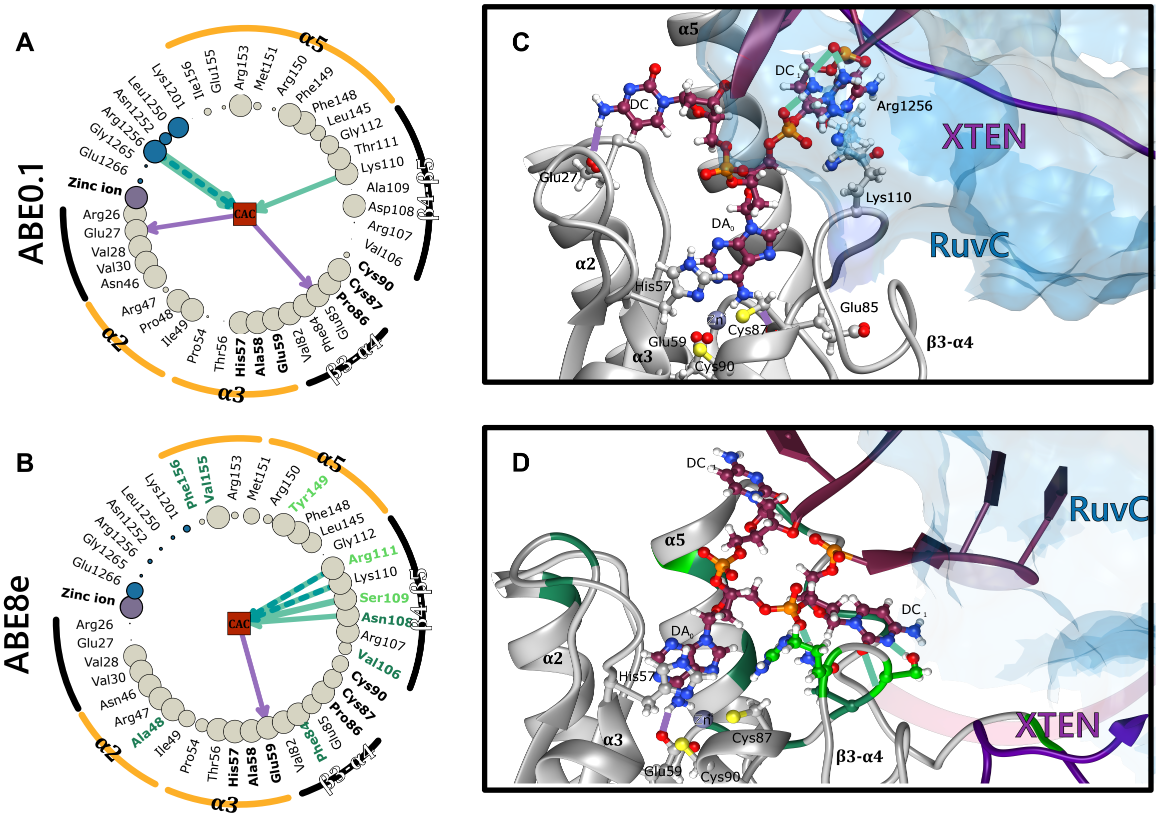


Figure 4.3: **Analyses of ABEs interaction with the target adenine and its flanking bases.** Asteroid plots for (A)ABE0.1 and (B) ABE8e complexes. (C) and (D) Details of changes in the molecular interaction between TadA*0.1 and TadA*8e with the target adenine and its flanking bases.

ssDNA makes with the TadA* domain of the ABE variants, we expanded the hydrogen bonding analysis beyond bases in the active site of TadA*(**Figure 4.3**). As expected the rigidity of the terminal α5 helix region of TadA*0.1 and the associated XTEN linker region, leads to extensive hydrogen bonding interaction with the bases upstream of the target adenine. Ten residues in the C-terminal helix of TadA*0.1 come in close proximity to the exposed ssDNA bases in the R-loop(**Figure 4.4**). However, this is not observed in the case of the TadA*8e, as the only interaction formed by the terminal α5 helix in this case is between Lys110 and the phosphate backbone of the +2 base.

These differences in the interaction with the exposed ssDNA, further elucidate the atomistic role played by the mutations in the C-terminal helix region of TadA*, which otherwise leverages this region of the enzyme to stabilize the tRNA hairpin loop of substrate tRNA.

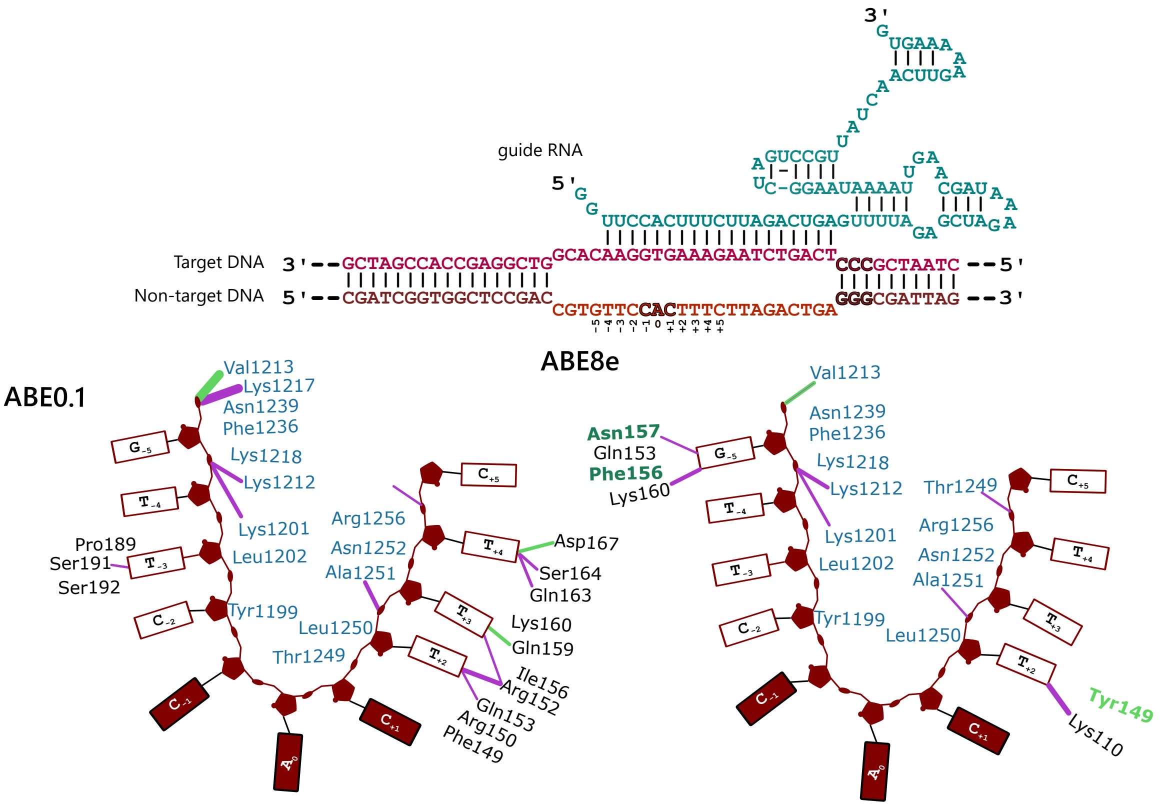


Figure 4.4: **Analyses of ABEs interaction with the bases in the exposed ssDNA.** Summary of ABE*-ssDNA interactions. The ssDNA represents exposed bases in the R-loop, which interact primarily with TadA* domain and the RuvC domain of Cas9.

4.4 Discussion

While traditional structure-based rational design of enzymes involves mutagenesis of the residues in the first or second interaction shell of substrate and active site pocket, the directed evolution of ABEs stands in contrast to this as most of the evolutionary mutations are situated >10 $\AA$ away from the target Adenine base. Although this number seems striking at first, but given that most residues of a protein actually reside on surface-exposed regions rather than the core/active site of the enzyme and as directed evolution can mutate all residues in an enzyme without any bias toward the surface/core regions, the abundance of remote mutations during directed evolution is a typical outcome. However, in the case of ABE's evolution, the concentration of mutations in the α5-helix region of the TadA*, which connects the C-terminal end of TadA* to the XTEN linker and subsequently the Cas9 enzyme is a consequence of not just the directed evolution scheme but also the architecture of the base editor enzymes.

Here we explored the structural and conformational significance of these remote mutations concentrated in the α5-helix region of the TadA* and found that such remote mutations helped the enzyme to diversify it deamination activity towards ssDNA substrates. Additionally, these mutations cause a kink in the terminal α5-helix which allows the enzyme to act on the target Adenine base without any constraints on the bases flanking the target base. This is critical for improving the therapeutic viability of the ABEs due to two reasons: firstly, it reduces the off target activity of the ABEs on RNA substrate, and, secondly, it allows ABEs to act on all target sites and not just the native 5′-TACG-3′.

Given the significance of these remote mutations in base editors, there is a clear need for methods which can help us go beyond minimalistic models of just the DNA-editing module bound to the ssDNA and generate predictive poses for the full length base editor enzymes with the prospective DNA-editing modules bound to the R-loop-Cas9 structure.

4.5 Methods

4.5.1 System Preparation

The MD simulations for all ABE variants was performed starting from the full-length cryo-EM structure (PDB ID: 6VPC[67]). The missing amino acid residues in TadA*8e[1-4, 160-167], missing XTEN linker[168-200], and Cas9[910-915, 967-972, 1104-1120, 1562-1565] and exposed ssDNA bases [31-38] were modeled using Modeller 10.1.[107] The cryo-EM coordinates were kept fixed, and 100 independent models were generated for the ABE-R-loop structure. The top 10 models were selected based on the lowest DOPE score and Z-score value and the final model was picked after thorough visual inspection of these ten models and ensuring that no loops were entangles or knotted in a physiologically irrelevant conformation or clashed with the rest of the resolved structure. Catalytic ion Mg^{+2} was added to the HNH domain and Ala840 residue of the Cas9 was mutated back to His residue. Waters of crystallization were added in from PDB ID: 4UN3. All titratable residues were protonated using the H++ server employing the default settings.[44, 99]

In order to prevent any simulation artifacts, especially due to unwanted interaction between the flexible XTEN linker and the PAM-distal DNA double helix, additional 10 base pairs of DNA was build using Chimera, based on the missing DNA sequence density in the original cryo-EM structure(NTS strand 5′-CGATCGGTGG-3′). At first, this DNA decamer was constructed in isolation in Chimera, followed by manual adjustments to place this DNA double helix close to the existing PAM-distal DNA sequence. The phosphate bonds were created manually between these DNA sequences to covalently link the missing decamer to the resolved DNA base pairs and the atom names as well as numbering was fixed in order to reflect the connectivity between the new DNA bases and the pre-existing ones.

The structure of ABE0.1 was modeled using steps similar to the ones listed above, except that the TadA*8e was swapped out with TadA*0.1 which was modeled using the predictions from Alphafold2.[53] This was done primarily because the X-ray structure of wtTadA (PDB ID: 1Z3A)[17] has no density for the terminal 10 amino acids of the α5-helix of the protein. We attempted to use Modeller[107] to predict the structure of these missing residues by the outcome was a highly disordered structure which is likely due to Modeller's inability to accurately predict the structure when its presented uncapped or singly-capped loops. After TadA*0.1 was swapped with TadA*8e in PDB ID: 6VPC[67] in Chimera,[42] the rest of the steps followed for the ABE8e structure modeling were repeated to construct the full-length ABE0.1 structure for simulations.

All systems were solvated in rectangular TIP3P waters box with a buffer length of 13.5Å(90,000 waters).[43] Additional Na^{+} ions were added to the system to maintain electroneutrality. The protein was represented using Amber ff14SB[45] and the RNA was represented using RNA.OL3 force field[100–102] and the DNA was represented using bsc1 parameters.[45–47] The Zinc metal-containing active site of TadA* was represented with custom force field parameters obtained using the MCPB.py approach at B3LYP/6-31G* level of theory[103] previously shown to be effectively represent the TadA* active site.[106] While the parameterization of Mg^{+2} ions has been deemed difficult in the previous simulations of similar CRISPR-Cas system,[108] we found that default parameters which match TIP3P model performed well for the systems studied here.[109]

4.5.2 MD simulations

All MD simulations were performed under periodic boundary conditions using the CUDA accelerated version of PMEMD implemented in Amber20 suite of programs.[48–50] The structures were relaxed using a combination of steepest descent and conjugate gradient minimization. During the first minimization phase, all atoms, except the waters, were restrained with a 300 kcal/molÅ^2 force constant. During the second and final minimization phase, all the restraints were removed and the system was allowed to freely minimize. The long-range electrostatics were cut-off at 12 Å.

This was followed by multi-step heating. During the first phase, the system was heated from 0-100K, with the non-water atoms held with a 100 kcal/molÅ^2 force constant. This was followed by a 100-200K heating ramp where only the backbone atoms of the non-water atoms were restrained with a 100 kcal/molÅ^2 force constant. Finally, all restraints were removed and the system was heated to the final temperature of 300K. The Langevin thermostat was employed in these NVT simulations with a collision frequency of 1 ps^{-1}

Finally, NPT equilibrations were performed for 2000 ns for all systems using a Brendsen barostat to maintain a 1 bar pressure. The hydrogen atom bond length was constrained by implementing the SHAKE algorithm. All MD simulations were propagated in time using the velocity Verlet with a time step of 2 fs. The initial 200 ns were discarded to compare on the equilibrate dynamics of the various ABE systems.

The CPPTRAJ module implemented within Amber21 was used to analyze all the MD trajectories.[28,29] The visualization of the MD trajectories was rendered using Chimera, and data were plotted using Matplotlib.[54]

Chapter 5

Conclusions and Future Perspectives

Adenine base editors have become invaluable tools in the fields of biotechnology, basic sciences, and medicine. As they quickly move from the bench to the clinic, we have taken a step back here to discuss the molecular underpinnings for DNA-editing activity of TadA that form the basis of this revolutionary technology. Through computational modelling, simulations, and bioinformatics analyses we have retrospectively analyzed the structural and functional implications of the amino acid mutations that led to the evolution of TadA into a successful base editor. As the base editing toolbox requires expansion, additional DNA-editing enzymes[110] and the plethora of RNA-editing enzymes[111] need to be explored as potential starting point for novel base editors. Hence, the ultimate question that lies at the heart of the research presented in this dissertation is, what single letter amino acid changes in an existing nucleobase modifying enzyme make it better at introducing single letter DNA base edits. Towards this goal, we have attempted to learn the principles of enzyme design using ABEs as a prototypical base editor.

While each new enzyme would follow a unique trajectory to become a programmable base editor, a few generalizable design rules have emerged from our analyses of the evolutionary trajectory of TadA which can be extended towards the rational design of novel base editors. First, from a engineering and design perspective, base editors are quite modular in their composition.

Under this design paradigm, the different modules are stitched together seamlessly to work in concert, and each component can be substituted for a similar effector. This has been demonstrated by varying the nature of the DNA-editing module (such as deaminases[10,11] which have been the primary focus of this thesis as well as glycosylases,[12–15] methyltransferases,[112–120] and even demethylases[112,119,121–125,125–132]) and the DNA targeting module (different Cas effectors), or by adding "accessory" modules to the architecture to influence DNA repair outcomes. Second, the individual modules can be sourced and designed or redesigned independently to optimize overall functionality. The motivations driving the engineering of the DNA-editing modules of base editors are to enhance their on-target editing efficiency, reduce their off-target DNA and/or RNA editing activity, or modulate their substrate scope and specificity. Third and most importantly, regardless of the protein engineering approaches employed, the mutations that have been identified have largely resided on structural motifs that interact with the substrate directly. This means that, regardless of the chemical properties of the enzyme and the method adopted for designing it, beneficial mutations tend to reside on active site/substrate binding loops of the DNA-editing module as well as the terminal motif connecting it to the DNA-binding module. While we focused here on the adenosine deaminase enzyme, TadA(**Figure 5.1**), this last design rule is immediately apparent in the mutational patterns seen in the various cytosine deaminases that form the basis of the CBEs(**Figure 5.2**) and even the double-stranded DNA deaminase DddA$_{tox}$ which forms the basis of mitochondrial CBEs(**Figure 5.3**).

Leveraging these general design strategies, in conjunction with computational approaches such as the ones explored in this dissertation could be the key to rational designing of novel base editors.

In Chapter 2, we carried out extensive computational simulations to decipher the importance and role of the critical first mutation that leads to changes in the substrate preference of TadA, from RNA to DNA. We observed that a single amino acid change is able to enhance the binding of the DNA substrates to TadA and thereby increase its A•T→G•C DNA-editing activity.

In Chapter 3, we showed that it is very hard to selectively suppress the native RNA-editing activity of TadA from its desirable DNA-editing activity, as the mutations that enhanced the binding of DNA to the TadA enzyme also showed increased binding with the RNA. In an attempt to find mutations that may abrogate TadA's off-target RNA editing activity, we developed a simple information entropy classification model. This model relies only upon the data acquired from sequences homologs of the starting-point enzyme, hence does not require time-intensive simulations or even a three-dimensional structure of the starting enzyme and can be used towards designing mutant libraries for high-throughput experimental screens.

In Chapter 4, we built upon the recently solved cryo-EM structure of the full-length ABE and latest innovations in the field of computational structure prediction to delve deeper into the role played by the remote mutations in TadA in context of the entire Cas9-R-loop complex. Using large scale simulations we showed that mutations in the C-terminal region of TadA change its conformation to fit ssDNA more readily. This enhanced ssDNA substrate engagement leads to broadening of the substrate specificity for DNA-editing by ABEs while also decreasing its RNA-editing. Hence, we speculate that although the native RNA-editing of the starting point enzyme is highly-correlated with its evolved DNA-editing activity, one way to suppress the former without compromising the latter is by introducing mutations to the protein that lead to conformational changes which prefer the R-loop's exposed ssDNA structure over the hairpin RNA structure.

Apart from the sequence-based bioinformatics analyses and the structure-based simulation methodologies explored in this dissertation for base editor design, there are many other computational approaches that are yet to be tested for their utility towards this highly complex design problem. Machine learning methods for enzyme design that have been grabbing a lot of attention recently is a noteworthy example(See references within[135]). While machine learning methods are as much of a black box as the method of directed evolution itself, it can be combined with sequence- and structure-based computational approaches such as the ones described throughout this thesis, to aid in rational design of base editors. I look forward to future research in the area of computational protein design where sequence-based machine learning and bioinformatics approaches are seemlessly combined with structure-based simulations to gain a predictive understanding of base editors and hence, make sense of the alphabet soup of life.

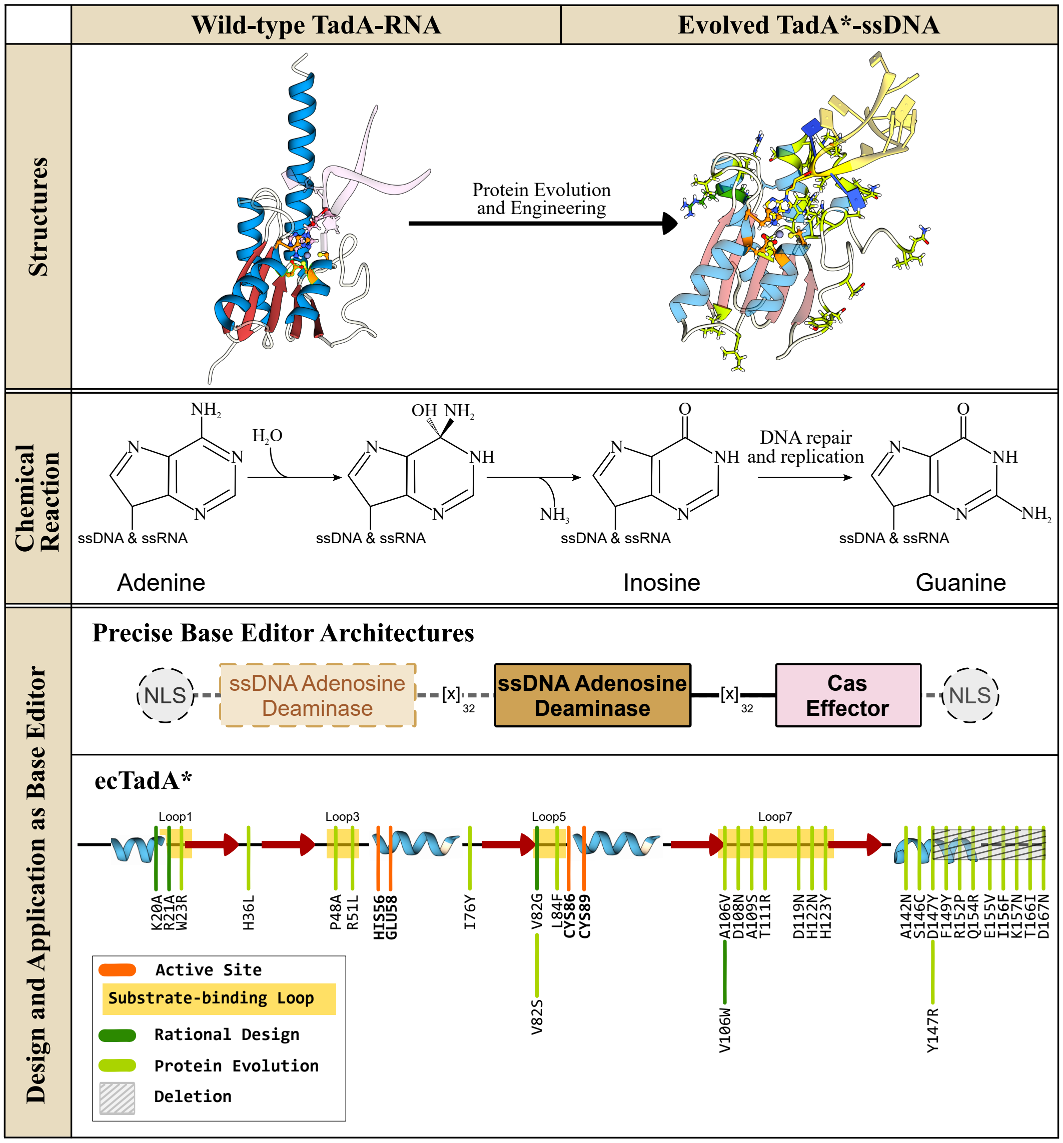


Figure 5.1: **Design of Adenine Base Editors.** ssDNA adenosine deaminase and its design and application in genomic DNA base editing. (Top) Structures of the wt-ecTadA enzyme bound to its substrate RNA, and the evolved TadA* bound to a ssDNA substrate. The conserved CDA fold is color-coded as follows: the β-sheet core is red and the peripheral α-helices are blue. The active site residues are shown in orange, with the mutations that have enhanced certain properties of the deaminase (when used as a base editor) shown in green. The wt-TadA-RNA structure is generated using a combination of Alphafold2 predictions[53] and PDB ID:2B3J,[30] while the TadA*-ssDNA structure is generated from PDB ID:6VPC.[67] (Middle) TadA* can hydrolytically deaminate adenines in ssDNA and ssRNA to yield an inosine, which is then processed into guanine via DNA replication and/or repair processes. Overall, this reaction gives rise to a A•T→G•C base pair conversion. (Bottom) Representative ABE architectures are shown, with the essential and non-essential components indicated with solid and dashed outlines, respectively. Secondary structure alignment of the TadA enzyme are shown, with an emphasis on the core CDA fold. Locations of the substrate-binding loops and active site residues are indicated, and key mutations discovered using either rational design or directed evolution approaches to enhance certain properties of the corresponding ABE are shown in dark and light green, respectively.

Figure 5.2: **Design of Cytosine Base Editors**. ssDNA cytidine deaminases and their design and applications in genomic DNA base editing. (Top) Structures of APOBEC/AID deaminases, which form the basis of the most extensively used CBEs. The conserved CDA fold is color-coded as follows: the β-sheet core is in red and peripheral α-helices are in blue. The active site residues are shown in orange, with mutations that have enhanced certain properties of these deaminases (when used as a base editor) shown in green. The APOBEC1, CDA, and AID structures are generated using Alphafold2[53] while APOBEC3A-ssDNA corresponds to PDB ID:5KEG.[133] (Middle) These enzymes can hydrolytically deaminate cytosines or 5-methylcytosines in ssDNA and RNA to yield a uridine or thymine bases, respectively. Overall, these reactions gives rise to C•G→T•A base pair conversions. (Bottom) Representative CBE architectures are shown, with essential and non-essential components indicated with solid and dashed outlines, respectively. Secondary structure alignments of APOBEC and AID deaminases are shown, with an emphasis on the similarity of their core CDA fold. Locations of the substrate-binding loops and active site residues are indicated, and key mutations discovered using either rational design or directed evolution approaches to enhance certain properties of the corresponding CBE are shown in dark and light green, respectively.

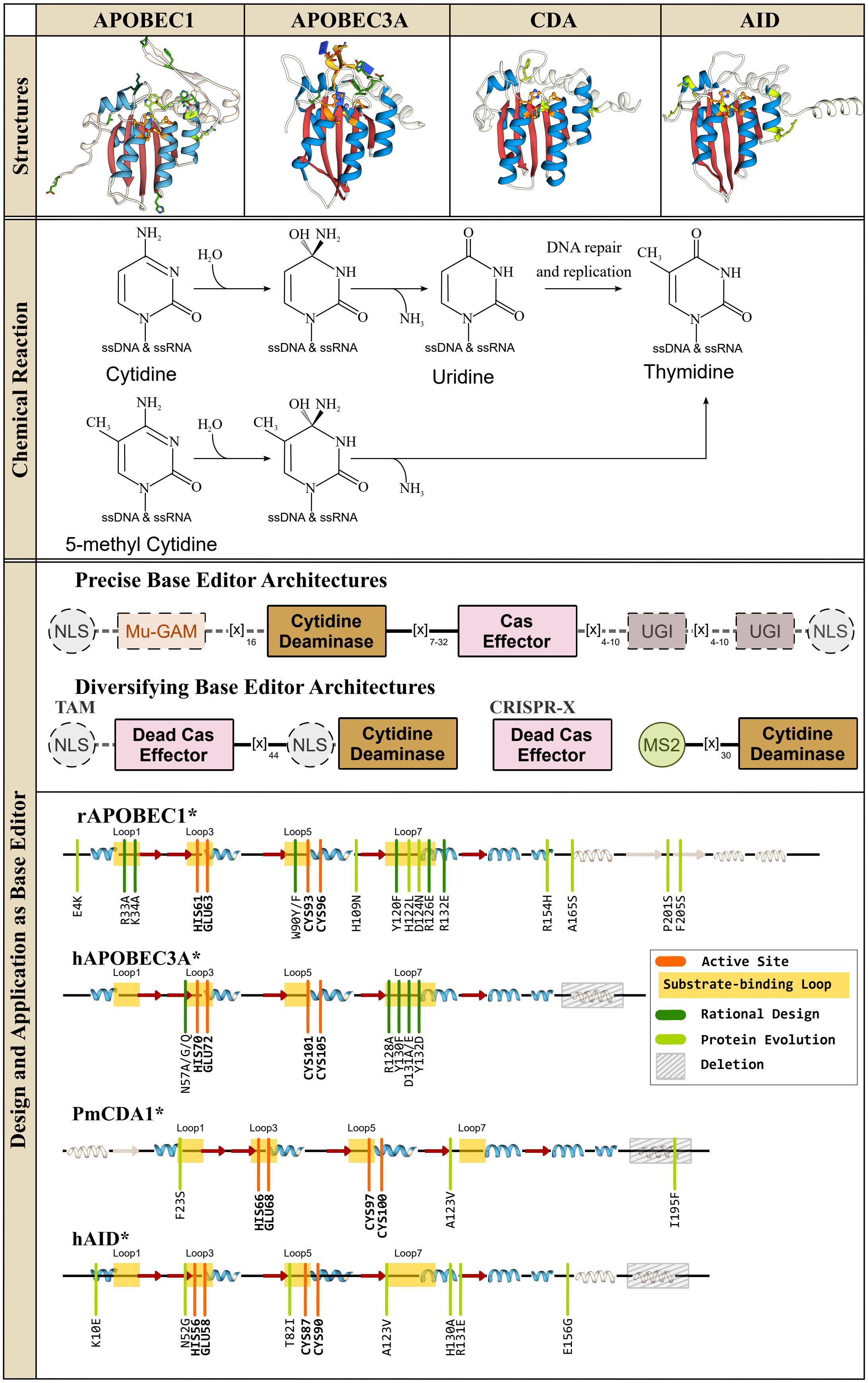
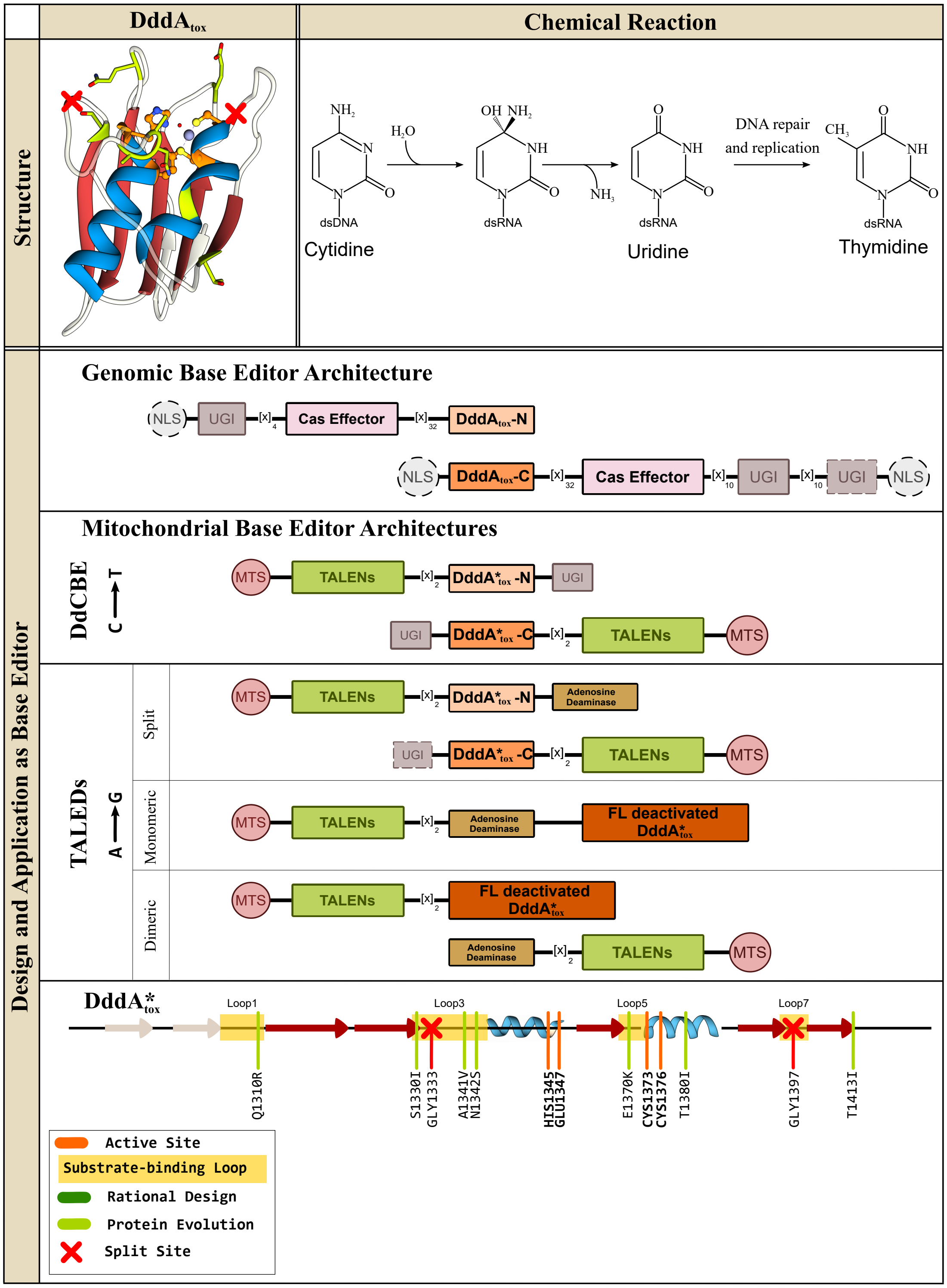

APOBEC1
APOBEC3A
CDA
AID
Structures
Chemical Reaction
NH2
N
O
N
ssDNA & ssRNA
Cytidine
H2O
OH NH2
NH
O
N
ssDNA & ssRNA
NH3
O
NH
O
N
ssDNA & ssRNA
Uridine
DNA repair
and replication
CH3
O
NH
O
N
ssDNA & ssRNA
Thymidine
CH3
NH2
N
O
N
ssDNA & ssRNA
5-methyl Cytidine
H2O
CH3
OH NH2
NH
O
N
ssDNA & ssRNA
NH3
Design and Application as Base Editor
Precise Base Editor Architectures
NLS
Mu-GAM
[x] 16
Cytidine Deaminase
[x] 7-32
Cas Effector
[x] 4-10
UGI
[x] 4-10
UGI
NLS
Diversifying Base Editor Architectures
TAM
NLS
Dead Cas Effector
[x] 44
NLS
Cytidine Deaminase
CRISPR-X
Dead Cas Effector
MS2
[x] 30
Cytidine Deaminase
rAPOBEC1*
Loop1
Loop3
Loop5
Loop7
E4K
R33A
K34A
HIS61
GLU63
W90Y/F
CYS93
CYS96
H109N
Y120F
H122L
D124N
R126E
R132E
R154H
A165S
P201S
F205S
hAPOBEC3A*
Loop1
Loop3
Loop5
Loop7
N57A/G/Q
HIS70
GLU72
CYS101
CYS105
R128A
Y130F
D131A/E
Y132D
Active Site
Substrate-binding Loop
Rational Design
Protein Evolution
Deletion
PmCDA1*
Loop1
Loop3
Loop5
Loop7
F23S
HIS66
GLU68
CYS97
CYS100
A123V
I195F
hAID*
Loop1
Loop3
Loop5
Loop7
K10E
N52G
HIS56
GLU58
T82I
CYS87
CYS90
A123V
H130A
R131E
E156G

Figure 5.3: **Design of Mitochondrial Cytosine Base Editors.** Deaminase toxin A (DddA$_{tox}$), a dsDNA cytidine deaminase, and its design and application in genomic and mitochondrial base editing. (Top left) Structure of the *B. cenocepacia* DddA$_{tox}$ enzyme (PDB ID: 6U08).[134] The conserved CDA fold is color-coded as follows: the β-sheet core is red and the peripheral α-helices are blue. The active site residues are shown in orange, with the mutations that have enhanced certain properties of the deaminase (when used as a base editor) shown in green. The split sites that were used to divide the enzyme into two inactive, non-toxic halves are shown as red crosses. (Top right) DddA$_{tox}$ can hydrolytically deaminate cytosines in dsDNA to yield a uridine, when is then processed into thymidine via DNA replication and/or repair processes. Overall, this reaction gives rise to a C•G→T•A base pair conversion. (Bottom) Representative DdCBE and DddA$_{tox}$-based ABEs (TALEDs) architectures are shown, with essential and non-essential components indicated with solid and dashed outlines, respectively. Secondary structure alignment of the DddA$_{tox}$ enzyme is shown, with an emphasis on the core CDA fold. Locations of the substrate-binding loops and active site residues are indicated, and key mutations discovered using directed evolution to enhance certain properties of the corresponding DdCBE are shown in light green and the split sites are indicated with red crosses.


DddA$_{tox}$
Chemical Reaction
Structure
NH_2
N
O
N
dsDNA
Cytidine
H_2O
OH NH_2
NH
dsRNA
NH_3
Uridine
DNA repair
and replication
CH_3
Thymidine
Design and Application as Base Editor
Genomic Base Editor Architecture
NLS
UGI
[x]$_4$
Cas Effector
[x]$_{32}$
DddA$_{tox}$-N
DddA$_{tox}$-C
[x]$_{10}$
Mitochondrial Base Editor Architectures
DdCBE
C → T
MTS
TALENs
[x]$_2$
DddA*$_{tox}$ -N
DddA*$_{tox}$ -C
TALEDs
A → G
Split
Monomeric
Dimeric
Adenosine Deaminase
FL deactivated DddA*$_{tox}$
DddA*$_{tox}$
Loop1
Loop3
Loop5
Loop7
Q1310R
S1330I
GLY1333
A1341V
N1342S
HIS1345
GLU1347
E1370K
CYS1373
CYS1376
T1380I
GLY1397
T1413I
Active Site
Substrate-binding Loop
Rational Design
Protein Evolution
Split Site

References

[1] Rodolphe Barrangou and Jennifer A Doudna. *Nat. Biotechnol.*, 34(9):933, 2016.

[2] Samuel H Sternberg and Jennifer A Doudna. Expanding the biologist's toolkit with crispr-cas9. *Molecular cell*, 58(4):568–574, 2015.

[3] David Benjamin Turitz Cox, Randall Jeffrey Platt, and Feng Zhang. Therapeutic genome editing: prospects and challenges. *Nature medicine*, 21(2):121, 2015.

[4] Holly A Rees and David R Liu. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nature reviews genetics*, 19(12):770–788, 2018.

[5] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

[6] Alexis C Komor, Ahmed H Badran, and David R Liu. Crispr-based technologies for the manipulation of eukaryotic genomes. *Cell*, 168(1-2):20–36, 2017.

[7] Elizabeth M Porto, Alexis C Komor, Ian M Slaymaker, and Gene W Yeo. Base editing: advances and therapeutic opportunities. *Nature Reviews Drug Discovery*, 19(12):839–859, 2020.

[8] Andrew V Anzalone, Luke W Koblan, and David R Liu. Genome editing with crispr–cas nucleases, base editors, transposases and prime editors. *Nature biotechnology*, 38(7):824–844, 2020.

[9] Tony P Huang, Gregory A Newby, and David R Liu. Precision genome editing using cytosine and adenine base editors in mammalian cells. *Nature Protocols*, 16(2):1089–1128, 2021.

[10] Alexis C Komor, Yongjoo B Kim, Michael S Packer, John A Zuris, and David R Liu. Programmable editing of a target base in genomic dna without double-stranded dna cleavage. *Nature*, 533(7603):420–424, 2016.

[11] Nicole M Gaudelli, Alexis C Komor, Holly A Rees, Michael S Packer, Ahmed H Badran, David I Bryson, and David R Liu. Programmable base editing of a• t to g• c in genomic dna without dna cleavage. *Nature*, 551(7681):464–471, 2017.

[12] Ibrahim C Kurt, Ronghao Zhou, Sowmya Iyer, Sara P Garcia, Bret R Miller, Lukas M Langner, Julian Grünewald, and J Keith Joung. Crispr c-to-g base editors for inducing targeted dna transversions in human cells. *Nature biotechnology*, 39(1):41–46, 2021.

[13] Liwei Chen, Jung Eun Park, Peter Paa, Priscilla D Rajakumar, Hong-Ting Prekop, Yi Ting Chew, Swathi N Manivannan, and Wei Leong Chew. Programmable c: G to g: C genome editing with crispr-cas9-directed base excision repair proteins. *Nature communications*, 12(1):1–7, 2021.

[14] Dongdong Zhao, Ju Li, Siwei Li, Xiuqing Xin, Muzi Hu, Marcus A Price, Susan J Rosser, Changhao Bi, and Xueli Zhang. Glycosylase base editors enable c-to-a and c-to-g base changes. *Nature biotechnology*, 39(1):35–40, 2021.

[15] Luke W Koblan, Mandana Arbab, Max W Shen, Jeffrey A Hussmann, Andrew V Anzalone, Jordan L Doman, Gregory A Newby, Dian Yang, Beverly Mok, Joseph M Replogle, et al. Efficient c• g-to-g• c base editors developed using crispri screens, target-library analysis, and machine learning. *Nature biotechnology*, 39(11):1414–1425, 2021.

[16] André P Gerber and Walter Keller. Rna editing by base deamination: more enzymes, more targets, new mysteries. *Trends in biochemical sciences*, 26(6):376–384, 2001.

[17] Jungwook Kim, Vladimir Malashkevich, Setu Roday, Michael Lisbin, Vern L Schramm, and Steven C Almo. Structural and kinetic characterization of escherichia coli tada, the wobble-specific trna deaminase. *Biochemistry*, 45(20):6407–6416, 2006.

[18] Hiroshi Nishimasu, F Ann Ran, Patrick D Hsu, Silvana Konermann, Soraya I Shehata, Naoshi Dohmae, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki. Crystal structure of cas9 in complex with guide rna and target dna. *Cell*, 156(5):935–949, 2014.

[19] Fuguo Jiang, David W Taylor, Janice S Chen, Jack E Kornfeld, Kaihong Zhou, Aubri J Thompson, Eva Nogales, and Jennifer A Doudna. Structures of a crispr-cas9 r-loop complex primed for dna cleavage. *Science*, 351(6275):867–871, 2016.

[20] Silvestro G Conticello. The aid/apobec family of nucleic acid mutators. *Genome biology*, 9(6):1–10, 2008.

[21] Reuben S Harris, Svend K Petersen-Mahrt, and Michael S Neuberger. Rna editing enzyme apobec1 and some of its homologs can act as dna mutators. *Molecular cell*, 10(5):1247–1253, 2002.

[22] Mary Anne T Rubio, Irena Pastar, Kirk W Gaston, Frank L Ragone, Christian J Janzen, George AM Cross, F Nina Papavasiliou, and Juan D Alfonzo. An adenosine-to-inosine

trna-editing enzyme that can perform c-to-u deamination of dna. *Proceedings of the National Academy of Sciences*, 104(19):7821–7826, 2007.

[23] Frances H Arnold. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

[24] Eric M Brustad and Frances H Arnold. Optimizing non-natural protein function with directed evolution. *Current opinion in chemical biology*, 15(2):201–210, 2011.

[25] Cong Huai, Gan Li, Ruijie Yao, Yingyi Zhang, Mi Cao, Liangliang Kong, Chenqiang Jia, Hui Yuan, Hongyan Chen, Daru Lu, et al. Structural insights into dna cleavage activation of crispr-cas9 system. *Nature communications*, 8(1):1–9, 2017.

[26] Shannon M Miller, Tina Wang, Peyton B Randolph, Mandana Arbab, Max W Shen, Tony P Huang, Zaneta Matuszek, Gregory A Newby, Holly A Rees, and David R Liu. Continuous evolution of spcas9 variants compatible with non-g pams. *Nature biotechnology*, 38(4):471–481, 2020.

[27] Edward N Baker and Roderick E Hubbard. Hydrogen bonding in globular proteins. *Progress in biophysics and molecular biology*, 44(2):97–179, 1984.

[28] Daniel R Roe and Thomas E Cheatham III. Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, 9(7):3084–3095, 2013.

[29] Daniel R Roe and Thomas E Cheatham III. Parallelization of cpptraj enables large scale analysis of molecular dynamics trajectory data, 2018.

[30] Heather C Losey, Alexander J Ruthenburg, and Gregory L Verdine. Crystal structure of staphylococcus aureus trna adenosine deaminase tada in complex with rna. *Nature structural & molecular biology*, 13(2):153–159, 2006.

[31] Holly A Rees, Christopher Wilson, Jordan L Doman, and David R Liu. Analysis and minimization of cellular rna editing by dna adenine base editors. *Science advances*, 5(5):eaax5717, 2019.

[32] Tina Wang, Ahmed H Badran, Tony P Huang, and David R Liu. Continuous directed evolution of proteins with improved soluble expression. *Nature chemical biology*, 14(10):972–980, 2018.

[33] Benjamin W Thuronyi, Luke W Koblan, Jonathan M Levy, Wei-Hsi Yeh, Christine Zheng, Gregory A Newby, Christopher Wilson, Mantu Bhaumik, Olga Shubina-Oleinik, Jeffrey R Holt, et al. Continuous evolution of base editors with expanded target compatibility and improved activity. *Nature biotechnology*, 37(9):1070–1079, 2019.

[34] Jason M Gehrke, Oliver Cervantes, M Kendell Clement, Yuxuan Wu, Jing Zeng, Daniel E Bauer, Luca Pinello, and J Keith Joung. An apobec3a-cas9 base editor with minimized bystander and off-target activities. *Nature biotechnology*, 36(10):977–982, 2018.

[35] Julian Grünewald, Ronghao Zhou, Sara P Garcia, Sowmya Iyer, Caleb A Lareau, Martin J Aryee, and J Keith Joung. Transcriptome-wide off-target rna editing induced by crispr-guided dna base editors. *Nature*, 569(7756):433–437, 2019.

[36] Ke Shi, Özlem Demir, Michael A Carpenter, Jeff Wagner, Kayo Kurahashi, Reuben S Harris, Rommie E Amaro, and Hideki Aihara. Conformational switch regulates the dna cytosine deaminase activity of human apobec3b. *Scientific reports*, 7(1):1–12, 2017.

[37] Ke Shi, Michael A Carpenter, Surajit Banerjee, Nadine M Shaban, Kayo Kurahashi, Daniel J Salamango, Jennifer L McCann, Gabriel J Starrett, Justin V Duffy, Özlem Demir, et al. Structural basis for targeted dna cytosine deamination and mutagenesis by apobec3a and apobec3b. *Nature structural & molecular biology*, 24(2):131–139, 2017.

[38] Lauren G Holden, Courtney Prochnow, Y Paul Chang, Ronda Bransteitter, Linda Chelico, Udayaditya Sen, Raymond C Stevens, Myron F Goodman, and Xiaojiang S Chen. Crystal structure of the anti-viral apobec3g catalytic domain and functional implications. *Nature*, 456(7218):121–124, 2008.

[39] Jason D Salter, Ryan P Bennett, and Harold C Smith. The apobec protein family: united by structure, divergent in function. *Trends in biochemical sciences*, 41(7):578–594, 2016.

[40] Silvestro G Conticello, Marc-Andre Langlois, and Michael S Neuberger. Insights into dna deaminases. *Nature structural & molecular biology*, 14(1):7–9, 2007.

[41] Julian Grünewald, Ronghao Zhou, Sowmya Iyer, Caleb A Lareau, Sara P Garcia, Martin J Aryee, and J Keith Joung. Crispr dna base editors with reduced rna off-target and self-editing activities. *Nature biotechnology*, 37(9):1041–1048, 2019.

[42] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[43] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[44] John C Gordon, Jonathan B Myers, Timothy Folta, Valia Shoja, Lenwood S Heath, and Alexey Onufriev. H++: a server for estimating p k as and adding missing hydrogens to macromolecules. *Nucleic acids research*, 33(suppl_2):W368–W371, 2005.

[45] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

[46] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, et al. Parmbsc1: a refined force field for dna simulations. *Nature methods*, 13(1):55–58, 2016.

[47] Romelia Salomon-Ferrer, Andreas W Gotz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. *Journal of chemical theory and computation*, 9(9):3878–3888, 2013.

[48] Romelia Salomon-Ferrer, David A Case, and Ross C Walker. An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210, 2013.

[49] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[50] DA Case, IY Ben-Shalom, SR Brozell, DS Cerutti, TE Cheatham III, VWD Cruzeiro, TA Darden, RE Duke, D Ghoreishi, MK Gilson, et al. Amber; university of california: San francisco, 2018. *Google Scholar There is no corresponding record for this reference*.

[51] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of computational chemistry*, 13(8):1011–1021, 1992.

[52] Fangqiang Zhu and Gerhard Hummer. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *Journal of computational chemistry*, 33(4):453–465, 2012.

[53] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[54] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[55] Jeannette Wolf, André P Gerber, and Walter Keller. tada, an essential trna-specific adenosine deaminase from escherichia coli. *EMBO J.*, 21(14):3841–3851, 2002.

[56] Yanting Zeng, Jianan Li, Guanglei Li, Shisheng Huang, Wenxia Yu, Yu Zhang, Dunjin Chen, Jia Chen, Jianqiao Liu, and Xingxu Huang. Correction of the marfan syndrome pathogenic fbn1 mutation by base editing in human cells and heterozygous embryos. *Mol. Ther.*, 26(11):2631–2637, 2018.

[57] Zhiquan Liu, Mao Chen, Siyu Chen, Jichao Deng, Yuning Song, Liangxue Lai, and Zhanjun Li. Highly efficient rna-guided base editing in rabbit. *Nat. Commun.*, 9(1):1–10, 2018.

[58] Chun-Qing Song, Tingting Jiang, Michelle Richter, Luke H Rhym, Luke W Koblan, Maria Paz Zafra, Emma M Schatoff, Jordan L Doman, Yueying Cao, Lukas E Dow, et al. Adenine base editing in an adult mouse model of tyrosinaemia. *Nat. Biomed. Eng.*, 4(1):125–130, 2020.

[59] Seuk-Min Ryu, Taeyoung Koo, Kyoungmi Kim, Kayeong Lim, Gayoung Baek, Sang-Tae Kim, Heon Seok Kim, Da-eun Kim, Hyunji Lee, Eugene Chung, et al. Adenine base editing in mouse embryos and an adult mouse model of duchenne muscular dystrophy. *Nature biotechnology*, 36(6):536–539, 2018.

[60] Kai Hua, Xiaoping Tao, Fengtong Yuan, Dong Wang, and Jian-Kang Zhu. Precise a· t to g· c base editing in the rice genome. *Mol. plant*, 11(4):627–630, 2018.

[61] Fang Yan, Yongjie Kuang, Bin Ren, Jingwen Wang, Dawei Zhang, Honghui Lin, Bing Yang, Xueping Zhou, and Huanbin Zhou. Highly efficient a· t to g· c base editing by cas9n-guided trna adenosine deaminase in rice. *Mol. plant*, 11(4):631–634, 2018.

[62] Shuai Jin, Yuan Zong, Qiang Gao, Zixu Zhu, Yanpeng Wang, Peng Qin, Chengzhi Liang, Daowen Wang, Jin-Long Qiu, Feng Zhang, et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science*, 364(6437):292–295, 2019.

[63] Erwei Zuo, Yidi Sun, Wu Wei, Tanglong Yuan, Wenqin Ying, Hao Sun, Liyun Yuan, Lars M Steinmetz, Yixue Li, and Hui Yang. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science*, 364(6437):289–292, 2019.

[64] Nicole M Gaudelli, Dieter K Lam, Holly A Rees, Noris M Solá-Esteves, Luis A Barrera, David A Born, Aaron Edwards, Jason M Gehrke, Seung-Joo Lee, Alexander J Liquori, et al. Directed evolution of adenine base editors with increased activity and therapeutic application. *Nature biotechnology*, 38(7):892–900, 2020.

[65] Michelle F Richter, Kevin T Zhao, Elliot Eton, Audrone Lapinaite, Gregory A Newby, Benjamin W Thuronyi, Christopher Wilson, Luke W Koblan, Jing Zeng, Daniel E Bauer, et al. Phage-assisted evolution of an adenine base editor with improved cas domain compatibility and activity. *Nature biotechnology*, 38(7):883–891, 2020.

[66] Changyang Zhou, Yidi Sun, Rui Yan, Yajing Liu, Erwei Zuo, Chan Gu, Linxiao Han, Yu Wei, Xinde Hu, Rong Zeng, et al. Off-target rna mutation induced by dna base editing and its elimination by mutagenesis. *Nature*, 571(7764):275–278, 2019.

[67] Audrone Lapinaite, Gavin J Knott, Cody M Palumbo, Enrique Lin-Shiao, Michelle F Richter, Kevin T Zhao, Peter A Beal, David R Liu, and Jennifer A Doudna. Dna capture by a crispr-cas9–guided adenine base editor. *Science*, 369(6503):566–571, 2020.

[68] Frances H Arnold. Unnatural selection: molecular sex for fun and profit. *Engineering and Science*, 62(1):40–50, 1999.

[69] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[70] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47(D1):D506–D515, 2019.

[71] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.*, 28(1):45–48, 2000.

[72] Javier F Chaparro-Riggers, Karen M Polizzi, and Andreas S Bommarius. Better library design: data-driven protein engineering. *Biotechnology Journal: Healthcare Nutrition Technology*, 2(2):180–191, 2007.

[73] Yi Yu, Thomas C Leete, David A Born, Lauren Young, Luis A Barrera, Seung-Joo Lee, Holly A Rees, Giuseppe Ciaramella, and Nicole M Gaudelli. Cytosine base editors with minimized unguided dna and rna off-target events and high on-target activity. *Nature communications*, 11(1):1–10, 2020.

[74] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. Ncbi blast: a better web interface. *Nucleic Acids Res.*, 36(suppl_2):W5–W9, 2008.

[75] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[76] Chris Sander and Reinhard Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Bioinf.*, 9(1):56–68, 1991.

[77] WS Valdar. Scoring residue conservation. *Proteins Struct. Funct. Bioinf.*, 48(2):227, 2002.

[78] John A Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[79] Christopher A Voigt, Stephen L Mayo, Frances H Arnold, and Zhen-Gang Wang. Computationally focusing the directed evolution of proteins. *Journal of Cellular Biochemistry*, 84(S37):58–63, 2001.

[80] Anthony A Fodor and Richard W Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2):211–221, 2004.

[81] Kartik L Rallapalli, Alexis C Komor, and Francesco Paesani. Computer simulations explain mutation-induced effects on the dna editing by adenine base editors. *Science advances*, 6(10):eaaz2309, 2020.

[82] SeHee Park and Peter A Beal. Off-target editing by crispr-guided dna base editors. *Biochemistry*, 58(36):3727–3734, 2019.

[83] Alan Grossfield. Wham: the weighted histogram analysis method.

[84] Ross C Walker, Michael F Crowley, and David A Case. The implementation of a fast and accurate qm/mm potential method in amber. *J. Comput. Chem.*, 29(7):1019–1031, 2008.

[85] Michael Gaus, Qiang Cui, and Marcus Elstner. Dftb3: extension of the self-consistent-charge density-functional tight-binding method (scc-dftb). *J. Chem. Theory Comput.*, 7(4):931–948, 2011.

[86] Xiya Lu, Michael Gaus, Marcus Elstner, and Qiang Cui. Parametrization of dftb3/3ob for magnesium and zinc for chemical and biological applications. *J. Phys. Chem. B*, 119(3):1062–1082, 2015.

[87] Marcus Elstner, Thomas Frauenheim, and Sándor Suhai. An approximate dft method for qm/mm simulations of biological structures and processes. *J. Mol. Struct. (THEOCHEM)*, 632(1-3):29–41, 2003.

[88] Dingguo Xu, Qiang Cui, and Hua Guo. Quantum mechanical/molecular mechanical studies of zinc hydrolases. *International Reviews in Physical Chemistry*, 33(1):1–41, 2014.

[89] Dhruva K Chakravorty, Bing Wang, Chul Won Lee, David P Giedroc, and Kenneth M Merz Jr. Simulations of allosteric motions in the zinc sensor czra. *Journal of the American Chemical Society*, 134(7):3367–3376, 2012.

[90] Adam Pecina, Susanta Haldar, Jindrich Fanfrlik, René Meier, Jan Rezac, Martin Lepsik, and Pavel Hobza. Sqm/cosmo scoring function at the dftb3-d3h4 level: Unique identification of native protein–ligand poses. *Journal of chemical information and modeling*, 57(2):127–132, 2017.

[91] Qin Xu and Hong Guo. Quantum mechanical/molecular mechanical molecular dynamics simulations of cytidine deaminase: From stabilization of transition state analogues to catalytic mechanisms. *The Journal of Physical Chemistry B*, 108(7):2477–2483, 2004.

[92] Qin Xu, Haobo Guo, Andrey Gorin, and Hong Guo. Stabilization of a transition-state analogue at the active site of yeast cytosine deaminase: Importance of proton transfers. *The Journal of Physical Chemistry B*, 111(23):6501–6506, 2007.

[93] Haobo Guo, Niny Rao, Qin Xu, and Hong Guo. Origin of tight binding of a near-perfect transition-state analogue by cytidine deaminase: implications for enzyme catalysis. *Journal of the American Chemical Society*, 127(9):3191–3197, 2005.

[94] Xin Zhang, Yuan Zhao, Honggao Yan, Zexing Cao, and Yirong Mo. Combined qm (dft)/mm molecular dynamics simulations of the deamination of cytosine by yeast cytosine deaminase (y cd). *J. Comput. Chem.*, 37(13):1163–1174, 2016.

[95] Bianca Manta, Frank M Raushel, and Fahmi Himo. Reaction mechanism of zinc-dependent cytosine deaminase from escherichia coli: A quantum-chemical study. *J. Phys. Chem. B*, 118(21):5644–5652, 2014.

[96] Stepan Sklenak, Lishan Yao, Robert I Cukier, and Honggao Yan. Catalytic mechanism of yeast cytosine deaminase: An oniom computational study. *Journal of the American Chemical Society*, 126(45):14879–14889, 2004.

[97] Asmita Sen, Vandana Gaded, Prabha Jayapal, Gopalan Rajaraman, and Ruchi Anand. Insights into the dual shuttle catalytic mechanism of guanine deaminase. *The Journal of Physical Chemistry B*, 125(31):8814–8826, 2021.

[98] Toshiaki Matsubara, Masashi Ishikura, and Misako Aida. A quantum chemical study of the catalysis for cytidine deaminase: contribution of the extra water molecule. *Journal of chemical information and modeling*, 46(3):1276–1285, 2006.

[99] Ramu Anandakrishnan, Boris Aguilar, and Alexey V Onufriev. H++ 3.0: automating p k prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.*, 40(W1):W537–W541, 2012.

[100] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiri Sponer, Thomas E Cheatham III, Charles A Laughton, and Modesto Orozco. Refinement of the amber force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.*, 92(11):3817–3829, 2007.

[101] Marie Zgarbová, Michal Otyepka, Jiri Sponer, Arnost Mladek, Pavel Banas, Thomas E Cheatham III, and Petr Jurecka. Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.*, 7(9):2886–2902, 2011.

[102] Pavel Banás, Daniel Hollas, Marie Zgarbová, Petr Jurecka, Modesto Orozco, Thomas E Cheatham III, Jirí Sponer, and Michal Otyepka. Performance of molecular mechanics force fields for rna simulations: stability of uucg and gnra hairpins. *J. Chem. Theory Comput.*, 6(12):3836–3849, 2010.

[103] Pengfei Li and Kenneth M Merz Jr. Mcpb.py: A python based metal center parameter builder, 2016.

[104] Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690, 1997.

[105] Jesse D Bloom, Michelle M Meyer, Peter Meinhold, Christopher R Otey, Derek MacMillan, and Frances H Arnold. Evolving strategies for enzyme engineering. *Current opinion in structural biology*, 15(4):447–452, 2005.

[106] Kartik L Rallapalli, Brodie L Ranzau, Kaushik R Ganapathy, Francesco Paesani, and Alexis C Komor. Combined theoretical, bioinformatic, and biochemical analyses of rna editing by adenine base editors. *The CRISPR Journal*, 5(2):294–310, 2022.

[107] Benjamin Webb and Andrej Sali. Protein structure modeling with modeller. In *Structural genomics*, pages 239–255. Springer, 2021.

[108] Giulia Palermo, Yinglong Miao, Ross C Walker, Martin Jinek, and J Andrew McCammon. Crispr-cas9 conformational activation as elucidated from enhanced molecular simulations. *Proceedings of the National Academy of Sciences*, 114(28):7260–7265, 2017.

[109] Pengfei Li and Kenneth M Merz Jr. Taking into account the ion-induced dipole interaction in the nonbonded model of ions. *Journal of chemical theory and computation*, 10(1):289–297, 2014.

[110] Ankur Jai Sood, Coby Viner, and Michael M Hoffman. Dnamod: the dna modification database. *Journal of cheminformatics*, 11(1):1–10, 2019.

[111] Pietro Boccaletto, Filip Stefaniak, Angana Ray, Andrea Cappannini, Sunandan Mukherjee, Elżbieta Purta, Małgorzata Kurkowska, Niloofar Shirvanizadeh, Eliana Destefanis, Paula Groza, et al. Modomics: a database of rna modification pathways. 2021 update. *Nucleic Acids Research*, 50(D1):D231–D235, 2022.

[112] X Shawn Liu, Hao Wu, Xiong Ji, Yonatan Stelzer, Xuebing Wu, Szymon Czauderna, Jian Shu, Daniel Dadon, Richard A Young, and Rudolf Jaenisch. Editing dna methylation in the mammalian genome. *Cell*, 167(1):233–247, 2016.

[113] Aleksandar Vojta, Paula Dobrinić, Vanja Tadić, Luka Bočkor, Petra Korać, Boris Julg, Marija Klasić, and Vlatka Zoldoš. Repurposing the crispr-cas9 system for targeted dna methylation. *Nucleic acids research*, 44(12):5615–5628, 2016.

[114] James I McDonald, Hamza Celik, Lisa E Rois, Gregory Fishberger, Tolison Fowler, Ryan Rees, Ashley Kramer, Andrew Martens, John R Edwards, and Grant A Challen. Reprogrammable crispr/cas9-based system for inducing site-specific dna methylation. *Biology open*, 5(6):866–874, 2016.

[115] Angelo Amabile, Alessandro Migliara, Paola Capasso, Mauro Biffi, Davide Cittaro, Luigi Naldini, and Angelo Lombardo. Inheritable silencing of endogenous genes by hit-and-run targeted epigenetic editing. *Cell*, 167(1):219–232, 2016.

[116] Christina Galonska, Jocelyn Charlton, Alexandra L Mattei, Julie Donaghey, Kendell Clement, Hongcang Gu, Arman W Mohammad, Elena K Stamenova, Davide Cacchiarelli, Sven Klages, et al. Genome-wide tracking of dcas9-methyltransferase footprints. *Nature communications*, 9(1):1–9, 2018.

[117] Peter Stepper, Goran Kungulovski, Renata Z Jurkowska, Tamir Chandra, Felix Krueger, Richard Reinhardt, Wolf Reik, Albert Jeltsch, and Tomasz P Jurkowski. Efficient targeted dna methylation with chimeric dcas9–dnmt3a–dnmt3l methyltransferase. *Nucleic acids research*, 45(4):1703–1713, 2017.

[118] Henriette O'Geen, Sofie L Bates, Sakereh S Carter, Karly A Nisson, Julian Halmai, Kyle D Fink, Suhn K Rhie, Peggy J Farnham, and David J Segal. Ezh2-dcas9 and krab-dcas9 enable engineering of epigenetic memory in a context-dependent manner. *Epigenetics & chromatin*, 12(1):1–20, 2019.

[119] James K Nuñez, Jin Chen, Greg C Pommier, J Zachery Cogan, Joseph M Replogle, Carmen Adriaens, Gokul N Ramadoss, Quanming Shi, King L Hung, Avi J Samelson, et al. Genome-wide programmable transcriptional memory by crispr-based epigenome editing. *Cell*, 184(9):2503–2519, 2021.

[120] Lin Lin, Yong Liu, Fengping Xu, Jinrong Huang, Tina Fuglsang Daugaard, Trine Skov Petersen, Bettina Hansen, Lingfei Ye, Qing Zhou, Fang Fang, et al. Genome-wide determination of on-target and off-target characteristics for rna-guided dna methylation by dcas9 methyltransferases. *Gigascience*, 7(3):giy011, 2018.

[121] Samrat Roy Choudhury, Yi Cui, Katarzyna Lubecka, Barbara Stefanska, and Joseph Irudayaraj. Crispr-dcas9 mediated tet1 targeting for selective dna demethylation at brca1 promoter. *Oncotarget*, 7(29):46545, 2016.

[122] Masahiro Okada, Mitsuhiro Kanamori, Kazue Someya, Hiroko Nakatsukasa, and Akihiko Yoshimura. Stabilization of foxp3 expression by crispr-dcas9-based epigenome editing in mouse primary t cells. *Epigenetics & chromatin*, 10(1):1–17, 2017.

[123] X Shawn Liu, Hao Wu, Marine Krzisch, Xuebing Wu, John Graef, Julien Muffat, Denes Hnisz, Charles H Li, Bingbing Yuan, Chuanyun Xu, et al. Rescue of fragile x syndrome neurons by dna methylation editing of the fmr1 gene. *Cell*, 172(5):979–992, 2018.

[124] Jeong Gu Kang, Jin Suk Park, Jeong-Heosn Ko, and Yong-Sam Kim. Regulation of gene expression by altered promoter methylation using a crispr/cas9-mediated epigenetic editing system. *Scientific Reports*, 9(1):1–12, 2019.

[125] Valentin Baumann, Maximilian Wiesbeck, Christopher T Breunig, Julia M Braun, Anna Köferle, Jovica Ninkovic, Magdalena Götz, and Stefan H Stricker. Targeted removal of epigenetic barriers during transcriptional reprogramming. *Nature communications*, 10(1):1–12, 2019.

[126] Goran Josipović, Vanja Tadić, Marija Klasić, Vladimir Zanki, Ivona Bečeheli, Felicia Chung, Akram Ghantous, Toma Keser, Josip Madunić, Maria Bošković, et al. Antagonistic and synergistic epigenetic modulation using orthologous crispr/dcas9-based modular system. *Nucleic acids research*, 47(18):9637–9657, 2019.

[127] Xingbo Xu, Xiaoying Tan, Björn Tampe, Tim Wilhelmi, Melanie S Hulshoff, Shoji Saito, Tobias Moser, Raghu Kalluri, Gerd Hasenfuss, Elisabeth M Zeisberg, et al. High-fidelity crispr/cas9-based gene-specific hydroxymethylation rescues gene expression and attenuates renal fibrosis. *Nature communications*, 9(1):1–15, 2018.

[128] Nicolas Marx, Clemens Grünwald-Gruber, Nina Bydlinski, Heena Dhiman, Ly Ngoc Nguyen, Gerald Klanert, and Nicole Borth. Crispr-based targeted epigenetic editing enables gene expression modulation of the silenced beta-galactoside alpha-2, 6-sialyltransferase 1 in cho cells. *Biotechnology journal*, 13(10):1700217, 2018.

[129] Nozomi Hanzawa, Koshi Hashimoto, Xunmei Yuan, Kenichi Kawahori, Kazutaka Tsujimoto, Miho Hamaguchi, Toshiya Tanaka, Yuya Nagaoka, Hiroshi Nishina, Sumiyo Morita, et al. Targeted dna demethylation of the fgf21 promoter by crispr/dcas9-mediated epigenome editing. *Scientific reports*, 10(1):1–14, 2020.

[130] Sumiyo Morita, Hirofumi Noguchi, Takuro Horii, Kazuhiko Nakabayashi, Mika Kimura, Kohji Okamura, Atsuhiko Sakai, Hideyuki Nakashima, Kenichiro Hata, Kinichi Nakashima, et al. Targeted dna demethylation in vivo using dcas9–peptide repeat and scfv–tet1 catalytic domain fusions. *Nature biotechnology*, 34(10):1060–1065, 2016.

[131] Xingxing Xu, Yonghui Tao, Xiaobo Gao, Lei Zhang, Xufang Li, Weiguo Zou, Kangcheng Ruan, Feng Wang, Guo-liang Xu, and Ronggui Hu. A crispr-based approach for targeted dna demethylation. *Cell discovery*, 2(1):1–12, 2016.

[132] Aziz Taghbalout, Menghan Du, Nathaniel Jillette, Wojciech Rosikiewicz, Abhijit Rath, Christopher D Heinen, Sheng Li, and Albert W Cheng. Enhanced crispr-based dna demethylation by casilio-me-mediated rna-guided coupling of methylcytosine oxidation and dna repair pathways. *Nature communications*, 10(1):1–12, 2019.

[133] Takahide Kouno, Tania V Silvas, Brendan J Hilbert, Shivender Shandilya, Markus F Bohn, Brian A Kelch, William E Royer, Mohan Somasundaran, Nese Kurt Yilmaz, Hiroshi Matsuo, et al. Crystal structure of apobec3a bound to single-stranded dna reveals structural basis for cytidine deamination and specificity. *Nature communications*, 8(1):1–8, 2017.

[134] Beverly Y Mok, Marcos H de Moraes, Jun Zeng, Dustin E Bosch, Anna V Kotrys, Aditya Raguram, FoSheng Hsu, Matthew C Radey, S Brook Peterson, Vamsi K Mootha, et al. A bacterial cytidine deaminase toxin enables crispr-free mitochondrial base editing. *Nature*, 583(7817):631–637, 2020.

[135] K.K. Yang. Papers on machine learning for proteins. https://github.com/yangkky/Machine-learning-for-proteins#reviews, 2022.