PREDICTIVE MODELING OF NOVEL MUTATIONS TO DNA-EDITING METALLOENZYMES AND DEVELOPMENT OF IMPROVED QM/MM METHODS

Mark Alan Hix

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2021

APPROVED:

G. Andrés Cisneros, Major Professor
Thomas Cundari, Committee Member
Sushama Dandekar, Committee Member
Rebecca Weber, Committee Member
Linda Chelico, Committee Member
Lee Slaughter, Chair of the Department of Chemistry
Pamela Padilla, Dean of the College of Science
Victor Prybutok, Dean of the Toulouse Graduate School

Hix, Mark Alan. *Predictive Modeling of Novel Mutations to DNA-Editing Metalloenzymes and Development of Improved QM/MM Methods*. Doctor of Philosophy (Chemistry), December 2021, 193 pp., 2 tables, 39 figures, 6 appendices, 336 numbered references.

Molecular dynamics simulations and QM/MM calculations can provide insights into the structure and function of enzymes as well as changes due to mutations of the protein sequence.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Alice Walker who has been of immeasurable support through the entirety of my graduate school career and with whom I look forward to many more years of happiness and scientific exploration. Your help, encouragement, and support over the past five years have been the most wonderful gift. I would like to thank my advisor, Andrés Cisneros, for always pushing me just a little further, and for giving me the tools to find the answers rather than just supplying the answers themselves. I would like to thank all of my labmates for the conversations about science and for the social stuff between the science. I will miss our group coffee runs. I would like to specifically thank Emmett Leddin, my office partner and friend, for keeping me focused on the goal, for helping me learn about other perspectives besides my own, and for somehow knowing exactly when a setback in science requires a juice box. I would like to acknowledge my family, who have provided endless encouragement to me, shared in my setbacks, and celebrated my successes. The pride you have all expressed as I have made this journey has helped push me forward. Finally, my friends outside academia, who have kept me grounded and provided endless hours of fun. I could not have acheived any of this without all of you.

This dissertation is dedicated to Cathryn Hecht Hix, my mother, who passed away in February 2018 during my second semester of graduate school. I know she'd be proud of what I've accomplished.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND REVIEW

The broader scope of my doctoral work is focused on metalloenzymes, specifically ones which catalyze DNA-editing reactions such as the cytidine-to-uracil mutation signature of APOBEC3 family enzymes or the DNA methylation reaction of the M.MpeI methyltransferase. I worked on method development with the use of evolutionarily conserved amino acids in proteins combined with electron localization to improve the selection of QM regions in QM/MM calculations. I also worked on a smaller inorganic complex and explored the effect of different substituent groups on the formation of low-barrier hydrogen bonds in tridentate pincer ligands.

The APOBEC3 enzymes are most well-known for their involvement in innate human immunity to HIV as well as their anti-cancer activity over the course of normal function. My first project covered APOBEC3H and the orientation of the ssDNA substrate it acts upon. There is no crystal structure of the enzyme with substrate bound, which thus required modeling of multiple possible substrate binding orientations and ultimately led to the determination of the likely orientation and helped to pinpoint the amino acids responsible for the recognition of the preferred 5'-TCG-3' motif.[1] From this project, I was able to continue investigating APOBEC3H and a known cancer-related mutation occurring at position 121. This mutation results in the formation of a large hydrogen bonding network across a large portion of the protein, which in turn causes a destabilization of the dimer interface as well as a disruption of the active site, subsequently leading to a loss of expression. I proposed a rescue mutation that would reverse these effects and restore expression and activity, and my prediction was validated by my experimental collaborators.[2]

APOBEC3C, another member of the APOBEC family, has several residues that differ between the human and rhesus variants. The human A3C is a catalytically active dimer *in vivo*, while the rhesus is a monomer and is not active against HIV. My project focused on the different residues between the two species and how they impact the dimer interface and

dynamic motion of the enzyme. We found that the residue at position 144 can impact not only the strength and stability of the interfacial hydrogen bonding network, disrupting the dimer interface, but also the behavior of loop 1, a region previously shown to be important to HIV restriction activity.[3]

The methyltransferase M.MpeI is a promiscuous enzyme, which normally uses *S*-adenosyl-L-methionine to methylate a cytidine as part of epigenetic regulation. A single mutation was reported that enabled M.MpeI to more easily use a non-standard carboxylated form of its usual substrate to produce a new modified nucleotide. I studied the different interactions in the active site between these two variants to gain better understanding of how the substrate preference was shifted, then applied these principles to the prediction of a second mutation that would shift favorability to the nonstandard residue over the standard. The manuscript for this project is currently in preparation.

I investigated the smaller inorganic metal complex of the tridentate pincer ligands discussed in Chapter 2.[4] Pincer ligands are chelating agents that tightly bind transition metals, conferring high thermal stability. They have also been investigated for their use in catalysis, especially with respect to C-H bond activation.

I developed an improved approach to the determination of QM regions in QM/MM calculations of enzymes.[5] This method helps to provide guidance in the selection of QM regions in order to obtain more accurate results without the inclusion of more atoms than necessary. Use of this method allows researchers to perform relatively inexpensive single point calculations to determine if amino acids or other molecules may be approximated in the MM region of a QM/MM simulation without significant loss of accuracy compared to an inclusion in the QM region. In that work, we demonstrated that this method, combined with evolutionary conservation of sequence and structure in enzymes, provides a novel method of QM region selection that is both inexpensive and accurate.

I worked to implement a Minimum Free Energy Path optimization protocol in the open source program LICHEM. Path optimizations to a potential energy surface do not account for the motion of the atoms in the surrounding environment to the reaction, and lack

the kinetic energy component of the reaction free energy calculation. This implementation uses molecular dynamics to obtain average forces over time on a reactive QM region in QM/MM calculations to account for this atomic motion and provide more accurate energy values for a path optimization.

These works together have been largely focused on enzyme dynamics and the development of new or improved methods by which metalloenzymes may be investigated, and have helped to produce models that successfully predicted useful mutations.

CHAPTER 2

CHARACTERIZING HYDROGEN-BOND INTERACTIONS IN PYRAZINETETRACARBOXAMIDE COMPLEXES: INSIGHTS FROM EXPERIMENTAL AND QUANTUM TOPOLOGICAL ANALYSES

2.1. Introduction

Tridentate pincers are known for aggressive binding of transition-metal ions.[6] Nonetheless, reports of 1,4-dimetalated pincers are rare.[7–10] The SCS diplatinum(II) pincer of Loeb and Shimizu[7] and palladium(II) and platinum-(II) complexes with NCN amine-based dipincers by van Koten et al.,[8] Weck et al.,[9] and Zhang and Lei[10] are representative of known dimetalated pincer complexes. While these complexes contain 1,4-dimetalated benzenes, to our knowledge, we were the first to report dimetalated pyrazinetetracarboxamides.[11] Of particular interest in our previously reported dipalladium(II) structure was the presence of two very short hydrogen bonds(HBs) between the 2,3- and 5,6-adjacent carboxamide oxygenatoms, at 2.430(5) Å. Similar phenomena were found fortransition-metal complexes with pyrazine-2,3-dicarboxamides[12–16] and attributed to a delocalized intermediate between amidate and iminolate tautomers.[13,16]

In an extension of a previous study with the tetraethyl-substituted dipincer H_4L_1, we introduced extended hydrophobic chains, tetrahexyl (H4L2), and hydrophilic chains, 2-hydroxyethyl ethyl ether (H_4L_3) (Figure 2.1), to explore the influence of the pendant chains on the solubilities. We were also curious as to whether the short O$\cdots$H+$\cdots$O HBs found for the palladium(II) complex of H_4L_1 would also be observed with the new ligands.

This chapter is presented in its entirety from Lohrman, J.; Vázquez-Montelongo, E. A.; Pramanik, S.; Day, V. W.; Hix, M. A.; Bowman-James, K.; Cisneros, G. A. Characterizing Hydrogen-Bond Interactions in Pyrazinetetracarboxamide Complexes: Insights from Experimental and Quantum Topological Analyses. *Inorganic Chemistry* **2018**, *57* (16), 9775–9778. https://doi.org/10.1021/acs.inorgchem.8b00627.

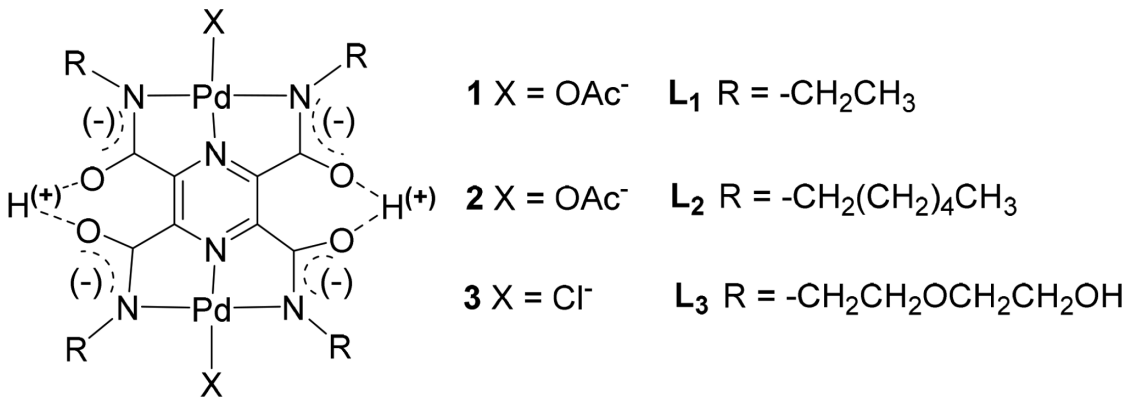


FIGURE 2.1. Pyrazinetetracarboxamide pincer complexes with ethyl ($\mathbf{L_1}$), hexyl ($\mathbf{L_2}$), and 2-hydroxyethyl ethyl ether ($\mathbf{L_3}$) substituents.

2.2. Computational Methods

To better characterize the nature of these HBs, here we present electron localization function (ELF),[17–19] noncovalent interaction (NCI),[20,21] and Bader's quantum theory of atoms in molecules (QTAIM)[22–27] analyses to investigate the nature of these interactions. It was previously shown that distinct HBs such as low-barrier hydrogen bonds (LBHBs) can be characterized by means of ELF analysis.[28,29] Additionally, combined ELF/NCI calculations are powerful tools that can provide insight into unconventional chemical bonding andweak NCIs in chemical and biological systems.[30–33]

2.3. Results

Pyrazine-2,3,5,6-tetracarboxamides are relatively easy to prepare from the tetrasubstituted methyl ester. The ethyl substituent gives the largest yield at almost 87% for the final step. The palladium(II) complexes were made as previously described for the acetate, with the exception that K_2PdCl_4 was used as the starting material for reaction with H_4L_3. Crystals of **1** and **3** suitable for X-ray diffraction were grown by slow evaporation from methanol, while crystals of **2** were obtained by vapor diffusion of ether into a CH_2Cl_2 solution of **2**.

All three palladium(II) complexes contain hydrogen atoms between each pair of adjacent amide carbonyl groups. The O---O distance varies only slightly, with O$\cdots$H+$\cdots$O distances equal to 2.430(5) Å for **1**, 2.427(4) and 2.426(4) Å for **2**, and 2.413(3) Å for **3**.

The O-H-O angles are not far from linear, 169(7), 174(6) (average of two), and 171(8)°, for **1**-**3**,respectively. The protons in each case were located in difference Fourier maps and refined isotropically. While it is true that the electron density between the two oxygen atoms was located and refined, the exact location of the hydrogen atom should not be taken as being highly accurate. Of the three complexes, in **3**, the protons were closest to the center between the two adjacent carbonyl groups, compared to **1** and **2** (Figure 2.2). The presence of this very short O···H+···O hydrogen bond is also manifested in solution, with sharp signals integrating to two protons at 19.43 ppm in $CDCl_3$ for **2** and 19.60 ppm in DMSO-d_6 for **3** (Figures S7 and S14).

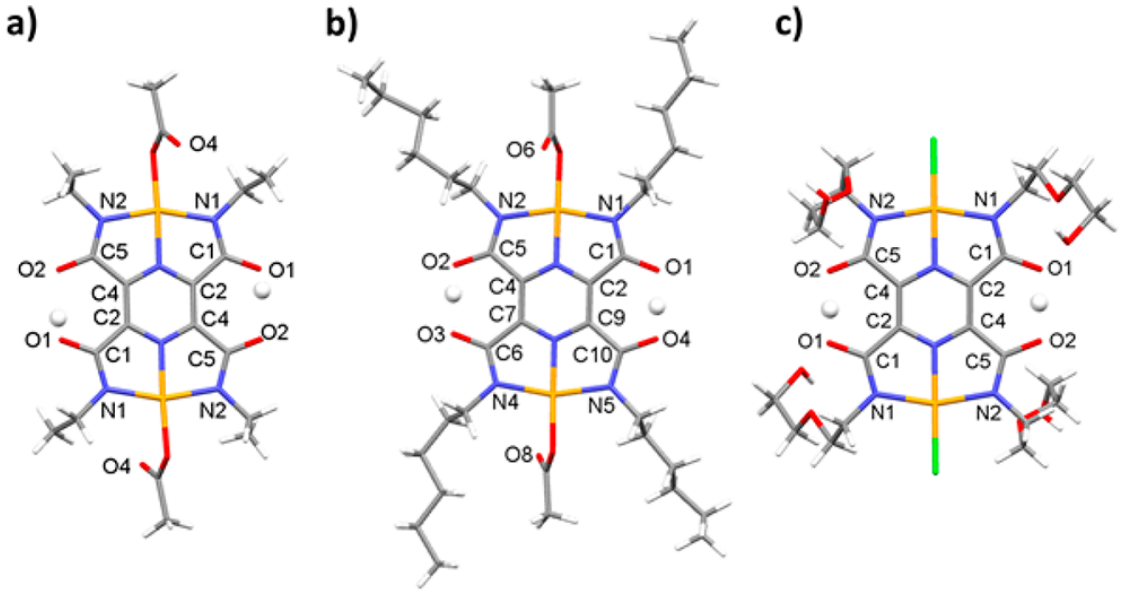


FIGURE 2.2. Overhead perspective views with selected labels of **1** (a), **2** (b), and **3** (c).

Selected bond lengths (Table S1) within the tautomeric regions are commensurate with approaching an intermediate between the amidate and iminolate tautomers. The amide C-O and C-N distances lie within narrow ranges, and average 1.280(5) and 1.300(5) Å, respectively, which is intermediate between single- and double-bond distances (Table S1). Other spectral measurements confirm the influence of the intermediate electron delocalization. In the solid-state IR spectra, the carbonyl stretches shift to lower energies for each of thecom-

plexes compared to the free ligand by about 50 cm^{-1} for **1** and **2**, and 60 cm^{-1} for **3** (Figures S7 and S14).

Intermolecular interactions between the nonbonded carboxylate oxygen atoms of the acetate and pyrazine π systems of neighboring complexes above and below in the stacked array are seen for **1** and **2**, with distances ranging from 2.669(5) to 2.853(5) Å (Figure 2.3a). The crystallographic packing for **3** is influenced by the presence of the extended hydrophilic chain(Figure 2.3b). Two opposing chains each capture a water molecule, which then forms very weak HB interactions (around 3.0 Å) with the adjacent hydroxyl group and oxygen atoms of two other neighboring complexes. The small differences in the O---O distances do not indicate significant intermolecular influence.

Although in our original report of these duplex pyrazine pincers we did preliminary density functional theory studies, those were primarily performed to investigate the possibility that **1** was a result of the reversion of the ligand to the iminol(ate) tautomeric form.[11] Those results indicated that **1** possessed a structure intermediate between the amidate and iminolate forms. In this contribution, we have performed quantum topological analyses to better understand the character of the observed HBs and, in particular, localization of the electrons based on the crystallographic findings.

ELF topological analyses of complexes **1**-**3** based on the geometries obtained from the crystal structures are shown in Figure 2.4 (combined ELF/NCI analyses are provided in Figures S24 and S25). Typically, hydrogen basins arising from ELF analyses are represented as covalent bonds (disynaptic basins), with the basin shared between the hydrogen atom and a heavy atom.[12,13] Strong interactions such as LBHBs are calculated by ELF to be valence basins associated with one single hydrogen atom (monosynaptic basin).[16]

In the case of complex **1**, the basin associated with the hydrogen atoms involved in the HB is seen to be shared (disynaptic basin) with the corresponding oxygen V(H1,O1)and V(H2,O1'). For complex **2**, one of the hydrogen atoms is associated with a monosynaptic basin [V(H1)], while the other is associated with a disynaptic one [V(H2,O4)]. Conversely, the basins in complex **3** are associated exclusively with the hydrogen atoms for both HBs,

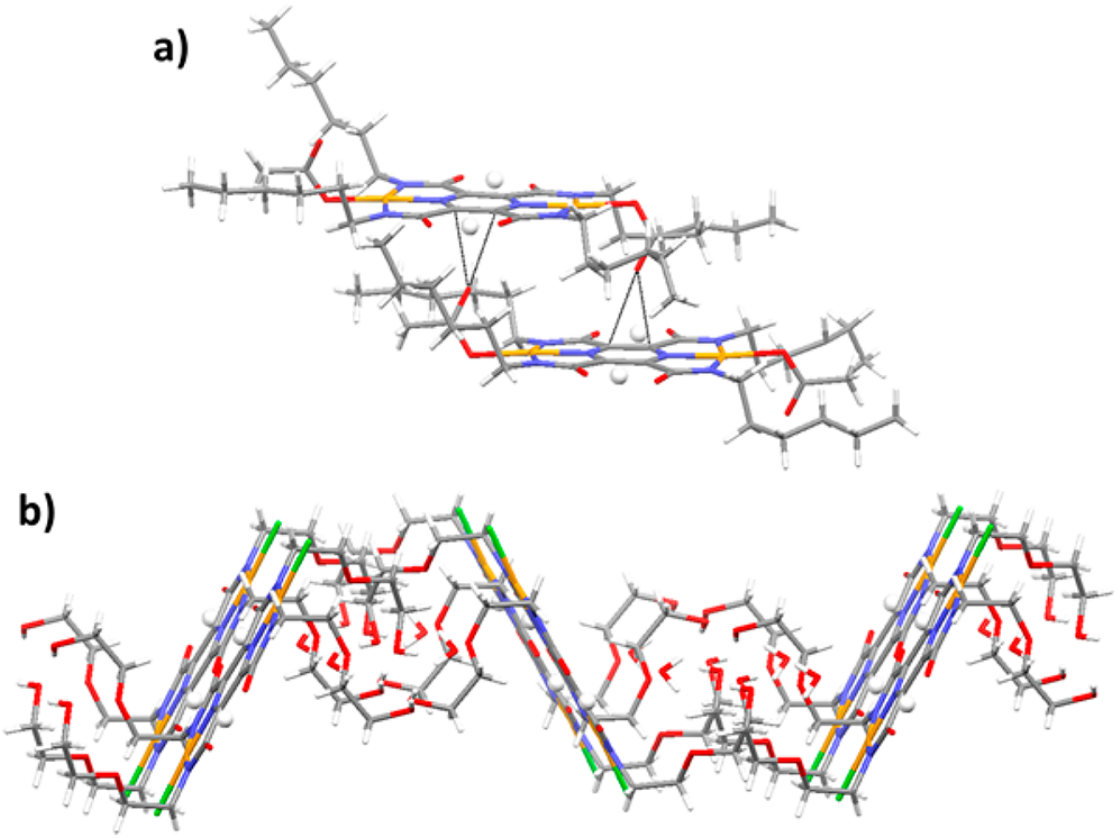


FIGURE 2.3. (a) Intermolecular stacking of the acetate complexes, as shown for **2** and (b) zigzag hydrophilic-hydrophobic chains of **3**.

V(H1) and V(H2). The combined ELF/NCI analysis (Figures S24 and S25) is consistent with the above results, except for complex **2**. For complex **1**, NCI analysis shows a dark blue surface for both HBs, indicating a strong HB for both interactions. In the case of **2**, disynaptic basin V(H2,O4) is predicted, although no strong interacting surface is observed between H2 and O4 from NCI analysis. For complex **3**, no surfaces are observed between the hydrogen atom and any of the oxygen atoms. Thus, the combined ELF/NCI results indicate that in **1** the interaction is basically short, but not LBHB; in **2**, one of the O$\cdots$H+$\cdots$O HB interactions is indicative of an LBHB, while the other is not; in **3**, with the shortest O---O distance, both interactions are indicative of LBHBs. It should be noted that crystallographic symmetry restrictions mandate that both HBs should be the same in **1** and **3** because no minimizations were performed.

The results of ELF population and distributed multipole analyses for selected basins are described below. Population and multipole analyses for complex **3** (Table A.4) show that the first and second polar moments for the lone-pair basins that are directly interacting with the hydrogen atom are significantly larger than the ones pointing away from the HB interaction region. This is an indication of the strong polarization on the oxygen atoms induced by the LBHB. In addition, population analysis shows that each of the oxygen atoms that share the hydrogen atom for complex **3** has a similar electron population (O2 = 4.03 and 1.62; O1 = 3.73 and 1.98); on the other hand,complex **1** (Table A.2) does not (O2 = 2.81 and 2.99; O1 = 3.86). This further suggests the presence of a covalent bond between H1 and O1 for complex **1**. For complex **2** (Table A.3), populations for the monosynaptic and disynaptic basins are similar to those for complexes **3** and **1**, respectively.

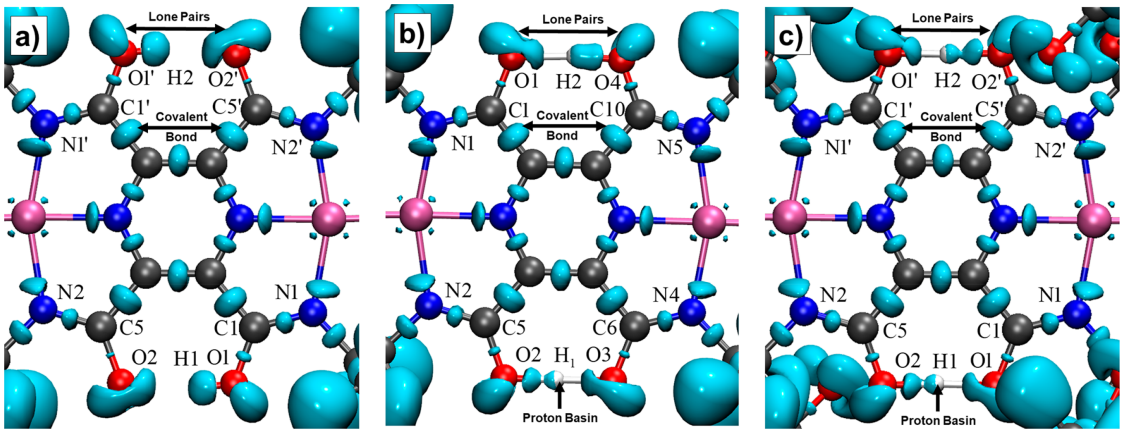


FIGURE 2.4. Closeup of ELF analysis for the two HB regions for complexes **1** (a), **2** (b), and **3** (c). Atoms involved in the HBs as well as lone pairs and covalent bond basins labeled for clarity. Isosurface cutoff for ELF = 0.85.

Complementary QTAIM analysis was carried out on complexes **1-3** (Tables S6-S8) to corroborate the results from ELF analysis. It has been previously reported that analysis of the density and Laplacian of the density at the bond critical points (BCPs) can be used to investigate different HBs.[34–37] Covalent bonds (shared interactions) possess negative values for $\nabla^2\rho(r)$ and small values for $\rho(r)$. Closed-shell interactions, like HBs, also have small $\rho(r)$ but positive $\nabla^2\rho(r)$ values. For complex **1** (Table S6), the $\nabla^2\rho(r)$ values for the BCPs

located between H1-O1 and H2-O1' were negative, while the $\nabla^2\rho(r)$ values for the BCPs at O1-H2 and O2'-H2 are positive. For complex **3** (Table S8), all of the $\nabla^2\rho(r)$ values were negative. For complex **2** (Table S7), the value of $\nabla^2\rho(r)$ for BCPs corresponding to H1-O2, O3-H1, and H2-O4 is negative, while the one corresponding to O1-H2 is positive. These results are consistent with ELF analysis and indicate that complex **1** exhibits a short HB, while **3** exhibits LBHBs; for complex **2**, one of the HB corresponds to an LBHB and the other to a short HB.

2.4. Conclusions

In summary, the duplex pincers provide an ideal platform for exploring the delicate interplay between the ligand strength and acid-base properties. The short O$\cdots$H+$\cdots$O HBs almost certainly possess high pK_a values, although with the exception of **3**, the complexes are not soluble in water. Similar short bonds have been observed in cyclopentadiene esters, where the driving force appears to be the formation of the highly stabl earomatic cyclopentadienides.[38] In that case, metal-ion coordination is not needed to trigger the HB formation. Furthermore, solution and solid-state experimental studies indicate that the intramolecular HBs may depend on other factors, including pincer pendant groups and fourth coordination site ancillary ligands. Computational findings indicate that localization or delocalization of the electrons within these very short HBs depends on very subtle differences in O---O separations, which, in turn, depends on the surrounding ligand field. We are continuing in our exploration of these HB phenomena.

CHAPTER 3

COMBINING EVOLUTIONARY CONSERVATION AND QUANTUM TOPOLOGICAL ANALYSES TO DETERMINE QM SUBSYSTEMS FOR BIOMOLECULAR QM/MM SIMULATIONS

3.1. Introduction

The atomistic investigation of enzymatic reaction mechanisms can provide useful insights for various applications.[39] A popular approach to investigate reaction mechanisms in enzymes is the hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) method. This approach combines two levels of theory, QM and MM, to allow the investigation of large systems by using a QM approach to represent a small region, at a minimum the atoms/molecules directly involved in the reaction, while the rest of the system is treated with a molecular mechanics (MM) method.[40–43]

The choice of molecular fragments to be included in the QM subsystem and approaches used to determine them is a topic of interest in the literature.[44–51] Several factors play a role in this selection, including (but not limited to) the overall computational time (wall–time) of the calculation, the QM level of theory (semi–empirical or *ab initio*)/basis set, the MM force field, the type of simulation to be performed (e.g. ground state reactivity, excited state reactivity, electronic excitation calculations, QM/MM MD), etc. The use of polarizable force fields has been shown to provide a better description of the MM environment, which can aid in the reduction of the number of atoms that need to be treated at the QM level.[52–58]

One of the ultimate goals of selecting the fragments for the QM subsystem is to achieve a balance between simulation accuracy and computational efficiency.[51] All QM/MM

This chapter is reprinted in its entirety from Hix, M. A.; Leddin, E. M.; Cisneros, G. A. Combining Evolutionary Conservation and Quantum Topological Analyses to Determine Quantum Mechanics Subsystems for Biomolecular Quantum Mechanics/Molecular Mechanics Simulations. *Journal of Chemical Theory and Computation* **2021**, *17* (7), 4524–4537. https://doi.org/10.1021/acs.jctc.1c00313.

calculations necessarily include the Minimum Active Region (MAR) in the QM subsystem, which we define here as the atoms or molecules *directly* involved in the chemical bond breaking/forming processes for the reaction under investigation

Intermolecular interactions between molecular fragments surrounding the MAR and the QM subsystem may or may not be approximated appropriately by the classical environment. In these cases, it is necessary to include additional fragments that may participate indirectly in the bond making and/or breaking processes. To this end, several methods have been proposed to determine the molecular fragments that need to be included in the QM region.

The QM fragment selection can be performed based on geometric considerations with respect to the MAR, such as including all atoms within a sphere or ellipsoid centered around the MAR,[45,48,49] or including fragments that comprise the 1st and 2nd shell of residues around the MAR.[32,59–68] Heuristic approaches that consider specific interactions of molecular fragments with the MAR have also been proposed. One such method uses covalent interactions, where residues which are accessible via a "covalent walk" from the MAR and lie within a specific radius are included.[69] Another approach relies on including fragments that form hydrogen bonds with molecules in the MAR.[70]

Other approaches for the determination of molecular fragments to be included in the QM subsystem rely on the calculation of electronic structure descriptors such as the charge deletion analysis[46] or the charge shift analysis.[71,72] The latter approach relies on the use of localized Fukui functions to determine the impact of the inclusion of additional residues beyond the MAR in enzyme catalysis. In this approach, single protein residues outside the MAR are included in the QM subsystem and changes to the localized Fukui function are evaluated. Subsequently, the residues that result in changes greater than a given threshold are included in the QM region.[71,72] A similar approach relies on the use use of QTAIM analysis to predict changes to enzymatic reaction barriers based on changing electric fields.[73]

In many cases, the inclusion of additional fragments in the QM subsystem results in

increased accuracy in the description of the system, however, this inclusion is commensurate with increased computational cost.[32,62,64,74,75]

Although there have been a variety of approaches developed for the selection of QM fragments, we are not aware of any procedure that considers protein evolutionary conservation for QM subsystem determination. Evolutionary conservation is a critical feature of enzyme families, and provides clues to catalytic activity and proficiency.[76,77] Therefore, it seems natural that the selection of QM fragments in simulations of enzymatic catalysis would benefit from the use of evolutionary conservation analysis. In this contribution we describe a method to select relevant fragments for the QM subsystem that relies on the use of protein sequence/structure evolutionary conservation to identify catalytically relevant residues; combined with a quantum interpretative technique to investigate the effects of specific fragments on the QM wavefunction.

Most proteins are classified into families based on structural and functional similarity, with variations on enzymatic activity, cellular localization, expression levels, and other characteristics.[78] These protein families often have structurally–conserved amino acid sequences across all members, with these residues frequently shown to be relevant to protein structure and/or function.[79–81] Pairs of unrelated proteins have also been shown to have conserved amino acids at critical positions in their sequences.[82] Some conserved residues may be of interest in investigations of enzymatic reactions as well, as previous studies have shown certain amino acids may be important for catalysis by interacting with reactant or transition states along the reaction path, without being directly involved in the breaking or forming of chemical bonds.[60,83,84]

As indicated above, the use of chemical descriptors based on QM can be used to investigate the effects of the environment on the QM wavefunction. The electron localization function (ELF) provides a way to investigate how regions of strong electron pairing are influenced in biological systems.[85] The ELF was initially proposed to calculate the range of electron pairing.[86,87] Since the ELF results in a continuous and differentiable scalar field, a topological analysis can be carried out (similar to QTAIM). The resulting scalar

and derivative fields can be divided into regions (basins). These basins have chemically intuitive interpretations including cores, lone pairs, bonds, etc. ELF has been shown to provide useful information for organic and enzymatic reactions.[88,89] Additionally, since the basins do not overlap and the electron population within each basin can be calculated, a multipolar expansion can be performed, to gain even further insights on these chemically relevant molecular regions.[90]

Here we present the use of an approach that combines protein sequence/structure evolutionary conservation and ELF analyses to determine the effects of various conserved residues and other molecular fragments on three enzyme systems: 4–oxalocrotonate (4OT), ten–eleven translocation–2 (TET2), and human DNA polymerase λ (Polλ). The remainder of the paper is as follows: the next section describes the computational methods for the evolutionary and ELF analysis, followed by a description of the computational approaches used to study the three enzymatic systems. Subsequently, the results of the conservation and ELF analysis on the two test systems, 4OT and TET2, are presented to show the applicability of ELF to the determination of appropriate fragments in the QM subsystem. Polλ provides a test system to determine the applicability of the method, followed by concluding remarks.

3.2. Computational Methods

Peptide sequence and structure alignments were performed using the T-Coffee Server in both Expresso mode and M-Coffee mode, and with Uniprot with CLUSTALO.[91–98] Conserved residues on the target systems from these multiple alignments were assessed by visual inspection to determine their proximity to the active site.

The three systems considered herein, 4OT, TET2, and Polλ, have been studied by QM/MM previously.[60–63,89,99,100] The details for the preparation of the simulation systems and computation of the individual paths of these three systems have been described previously. Briefly, all systems correspond to the wild type structures. In every case, the structures have been checked with PROPKA and MolProbity to assign ionizable residue protonation states and check side–chain rotamers. All systems are solvated in TIP3P water, neutralized with the corresponding number of required counterions and subjected to molec-

ular dynamics. For 4OT, the structure is based on pdbid: 1BJP,[101] the QM subsystem includes the N-terminal Pro (P1), 2–oxo–4–hexenedioate and (in some cases) R39" or R61' (see below).[60,99]

For TET2, the reference system, based on pdbid 4NM6,[102] is the same as the one used in the original calculations.[62] Here the QM subsystem includes the ferryl intermediate and all 1st–shell ligands including coordinating residues (H1382, D1384, H1881), succinate, water, 5–hydroxymethyl–Cytosine (substrate), 2nd–shell water, T1372 and Y1902. The two systems for the investigation of the role of different molecular fragments in the active site do not include either the 2nd–shell water or T1372/Y1902 in the QM subsystem (see below). All TET2 systems were calculated in the intermediate spin quintet state with Fe in the +3 oxidation state ferromagnetically coupled to the oxygen atom.[62]

In the case of Polλ the system is based on pdbid: 2PFO,[103,104] the QM subsystem comprises the two known (catalytic and nucleotide–binding) Mg^{2+}, 1st–coordination shell residues (D427, D429, D490), templating base, incoming nucleotide, cation–coordinating waters, a proposed 3rd Mg(II) cation with coordinating waters, and (in some cases) R386, R420, and K472 (see below).[63,89,100] QM/MM calculations (single point or optimization) were done at the B3LYP/6–31+G(d), ωB97X–D/6–311G(d,p), and B3LYP/6–31G(d) for 4OT, TET2 and Polλ, respectively with AMBER parm99 for the MM environment.[105] All calculations were performed with LICHEM interfacing Gaussian16 and TINKER.[106–109]

ELF analysis was performed with the TopMOD package to determine basin populations at selected basins involved with the specific reactions.[110] For several calculations, the Mod_wfn utility within the TopMOD package was also used to generate minimal wavefunctions only on the reactive atoms. Each system was subdivided into 200 grid points along each axis for a total of 8,120,601 segments and calculated using very high accuracy mode. Multiwfn was used for wavefunction analysis of the reactive atoms.[111] ELF heat maps were generated on a 200x200 grid using a plane defined by three reactive atoms. Close-ups of the heat map for Polλ were generated using a 500x500 grid. VMD and UCSF Chimera were used for visualizing the iso–surface values and creating images.[112,113]

3.3. Results

3.3.1 System Selection

The ultimate goal of the combined protein sequence/structure evolution and ELF analyses procedure is to determine whether this procedure can provide insights on whether a particular residue or molecular fragment should be included in the QM subsystem, or if it may be approximated by the MM potential. To this end, we have chosen systems that have previously been investigated via QM/MM and where specific molecular fragments have been shown to be appropriately represented by the QM or MM subsystems.

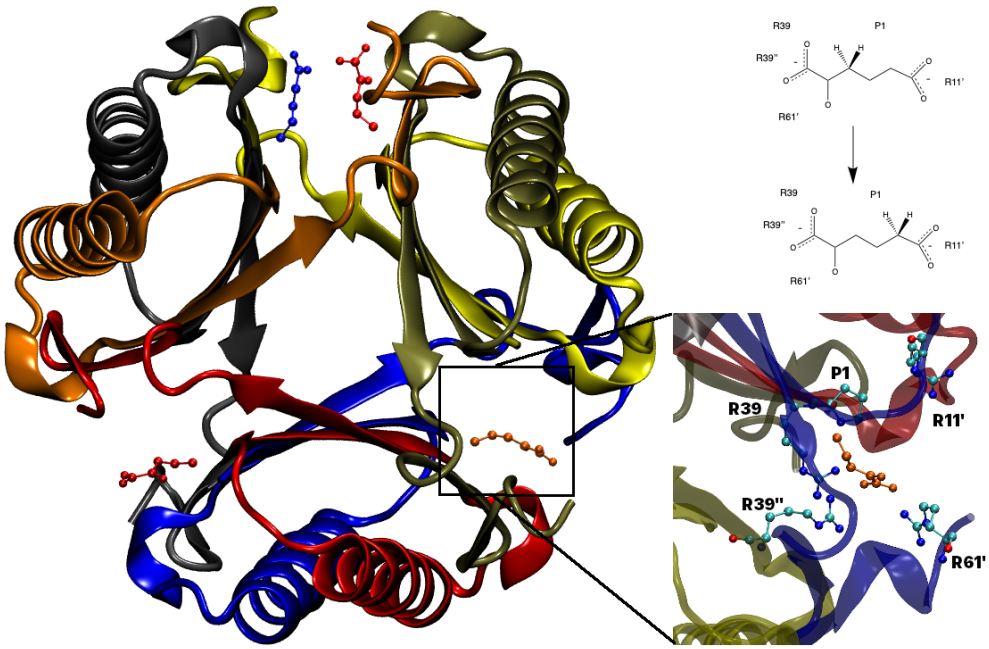


FIGURE 3.1. 4OT, clockwise from left: 4OT structure (pdb id: 5TIG),[114] scheme for the reaction mechanism catalyzed by 4OT, and close up of active site with conserved and active site residues.

4-oxalocrotonate tautomerase (4OT) catalyzes the tautomerization or 2–oxo–4–hexenedioate to 2–oxo–3–hexenedioate. The structure of 4OT is a so–called trimer of dimers comprised of 6 active sites (Figure 3.1), and has been shown to exhibit "half–of–the–sites" reactivity.[115] Each active site includes residues from three different monomers, thus, the residues are designated as unprimed, single–primed, or double–primed to denote that the

residues belong to different monomers (see Figure 3.1).[116] The reaction is carried out by an N–terminal P1 acting as a general base by abstracting the proton from the substrate, and proceeds in two steps (Figure 3.1).[60,61,99,117,118] Several charged residues are located in the active site including R11', R39", and R61' (Figure 3.1). These residues are highly conserved (as well as most of the rest of the monomer sequence) across homologs of 4OT, including the α/β heterohexamer 4OT (Figure 3.4a-b).[119]

The original QM/MM simulations showed that no general acid is required in the reaction,[60,99] subsequently confirmed experimentally by Metanis *et al.*;[120] instead, R39" stabilizes the TS1, intermediate, and TS2 structures by electrostatic interactions. QM/MM calculations have also shown that it is necessary to consider the full hexameric structure in order to provide the correct electrostatic environment for the active site under consideration.[61]

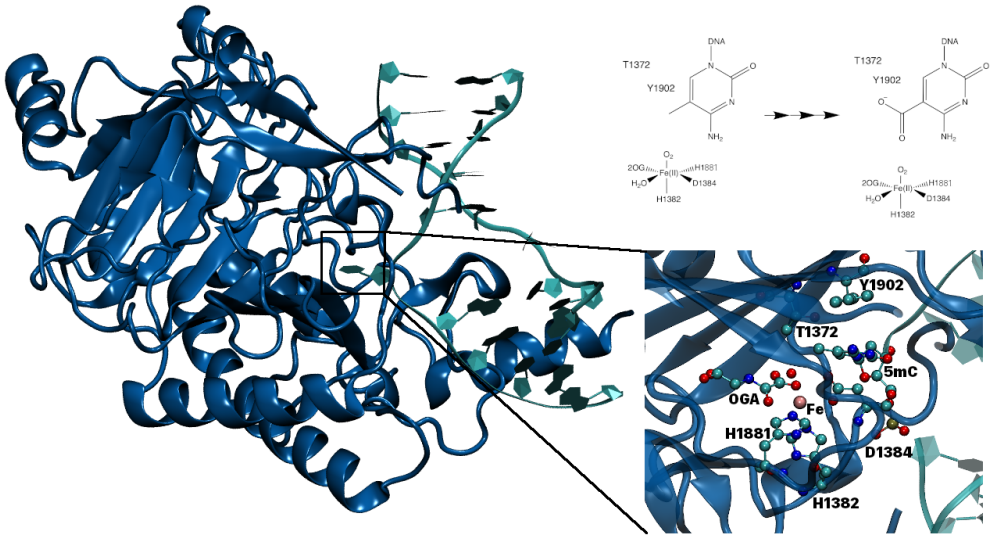


FIGURE 3.2. TET2, clockwise from left: TET2 structure (pdb id: 4NM6), scheme for the reaction mechanism catalyzed by TET2, and close up of active site with conserved and active site residues.

TET2 is part of the TET family, itself part of the Fe/α–ketoglutarate (αkg) super–family that catalyzes the sequential oxidation of 5–methylcytosine (5mC) $\rightarrow$ 5–hydroxymethylcytosine (5hmC) $\rightarrow$ 5–formylcytosine (5fC) $\rightarrow$ 5–carboxycytosine (5caC) (Figure

3.2).[75] A combined experimental and theoretical study showed that an active site scaffold formed by two highly conserved residues (Figure 3.4), T1372 and Y1902, helps to position the substrate in the active site to allow the oxidation by a ferryl (Fe(IV)=O) intermediate to proceed.[75] Subsequent QM/MM simulations showed that the rate–limiting step for the oxidation of 5hmC to 5fC proceeds through a ferryl intermediate (charateristic of Fe/αkg family enzymes) via a hydrogen atom abstraction from the hydroxyl of 5hmC, while a T1372A variant produces a change in orientation of 5hmC in the active site, resulting in a higher barrier.[62]

Additionally, the results from these QM/MM simulations suggest that a water molecule in the "2nd–shell" of the ferryl plays a role in the H atom abstraction (Figure 3.2).[62] Conserved water molecules in enzymes are known to play important roles in catalysis.[121,122] The second–shell ligand effect on ferryl–based catalysis of C–H bond breaking (activation) has been reported previously in other Fe/αkg family enzymes like AlkB,[123] inorganic complexes,[124] and Fe–based metal–organic frameworks.[125]

Polλ belongs to the X–family of polymerases and is involved in fixing double–stranded breaks in the non–homologous end–joining (NHEJ) pathway by catalyzing the nucleotide addition that fills small 1–2 residue gaps.[126] The reaction mechanism of Polλ proceeds via the deprotonation of the O3' on the primer base and subsequent nucleophilic attack on the α–phosphate of the incoming nucleotide with concomitant formation of pyrophosphate (Figure 3.3.[63] There are five residues conserved across human DNA polymerases in a second-shell around the active site (Figure 3.4).[63,100,127] Of these five, two are conserved across the X–family — for Polλ, these correspond to K472 and R488.

The presence of two metals in the active site, termed catalytic and nucleotide–binding, has been extensively supported in DNA polymerases.[63,128–131] Recently, a third divalent cation has been observed in the active sites of Polη and Polβ (another X–family polymerase).[132–134] Combined QM/MM simulations and structural/biochemical studies on the role of this third cation in the reaction mechanism of Polβ,[135–137] and QM/MM simulations on the reaction mechanism of Polη with a third cation,[138] indicate that this cation serves to

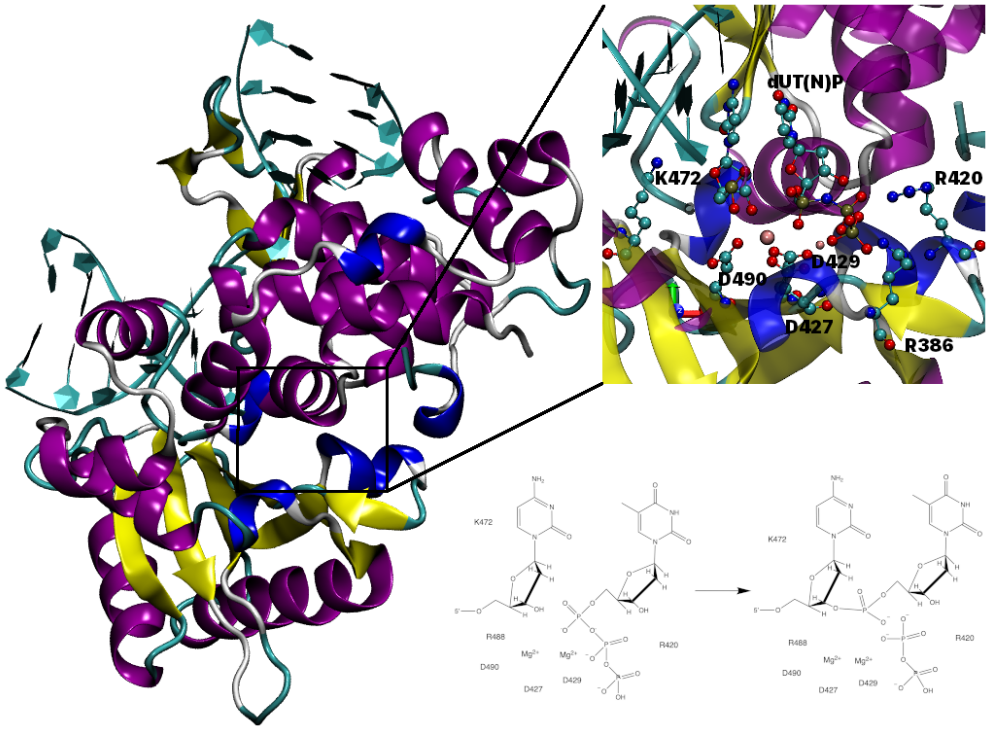


FIGURE 3.3. Polλ, counterclockwise from left: Polλ structure (pdbid: 2PFO), scheme for the reaction mechanism catalyzed by Polλ, and close up of active site with conserved and active site residues.

stabilize the product structure and avoid the reverse (pyrophosphorolysis) reaction. Prior QM/MM simulations on reactant and product structures of Polλ including a third cation in the active site found that the third metal resulted in a negative reaction energy,[100,106] compared with a positive reaction energy with only two cations.[63]

We have chosen these systems to test and validate our combined procedure given the role of different conserved residues and/or conserved waters in these systems. 4OT has been selected because it has been previously shown that a minimal representation of the active site is sufficient for an accurate QM/MM evaluation of the reaction path. Conversely, for TET2 it is important to include a water molecule in the QM subsystemthat is located in the 2nd–coordination shell (2nd–shell), and not directly involved in the making/breaking of covalent bonds or directly coordinating the metal center.

Based on this, we have performed a series of single point calculations on previously

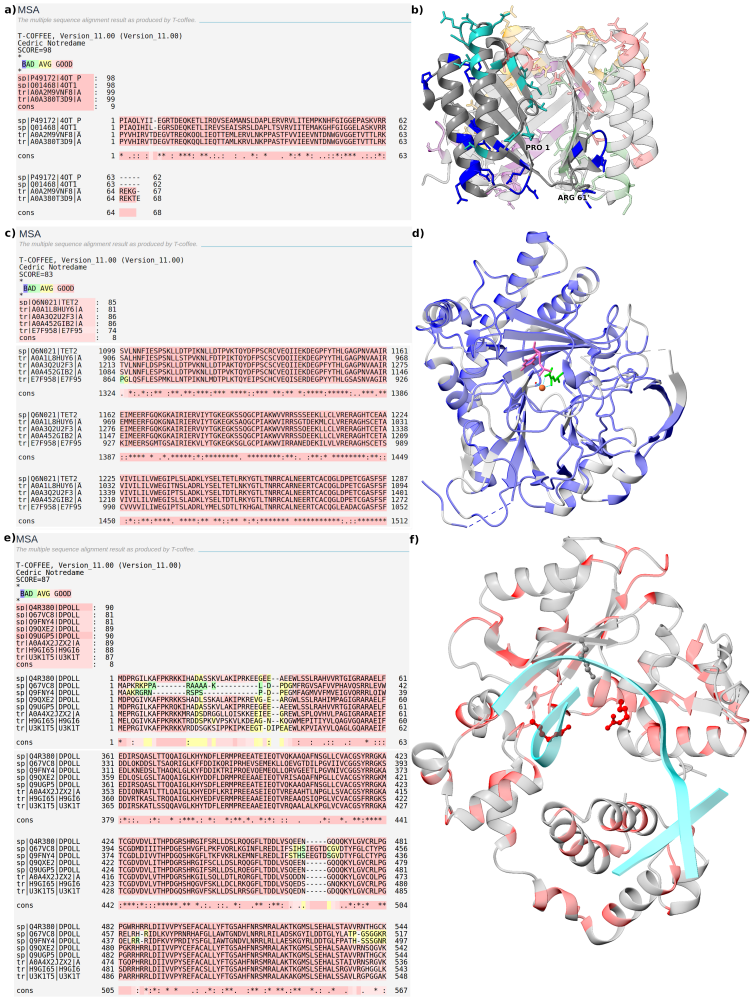


FIGURE 3.4. **a)** Sequence/structure alignment of 4OT homologs from four species of pseudomonas. **b)** Evolutionarily conserved residues on the 4OT hexamer structure with separate colors for each monomer subunit. **c)** Partial sequence/structure alignment of TET2 from multiple classes in the kingdom animalia. For complete alignment, see SI. **d)** Evolutionarily conserved residues of TET2 (blue) with substrate (pink) and cofactor (green) shown for reference, shown on the structure used in the previously reported study. **e)** Partial sequence/structure alignment of Polλ from several plant and animal species. **f)** Evolutionarily conserved residues shown in red, with residues tested in this study shown as ball-and-stick models.

optimized structures of 4OT and TET2 with different residues/molecular fragments in the QM subsystem (see below) to obtain the required wavefunctions for the ELF calculations to investigate the effects of QM or MM representations of critical residues/molecules around the active site. Subsequently, we have applied the combined procedure on the reactant structure of Polλ, including the putative third cation, to determine which residues to include in the QM subsystem followed by a full reaction path optimization with the third cation in the active site.

3.3.2 QM Subsystem Investigation for 4OT and TET2

The catalytic proficiency of enzymes is given in large part by electrostatic stabilization provided by the protein environment. Therefore, if a residue or molecular fragment is accurately approximated by the MM environment, the electronic charge distribution for the QM subsystem should be similar whether that residue or fragment is included in the QM subsytem or located in the MM environment. This should hold for both the reactant, and transition state (TS) structures.

We have performed ELF analysis on previously optimized reactant and TS structures for 4OT and TET2 to test this hypothesis. These two systems have been chosen to test this approach because they provide examples of systems where nearby molecular fragments have been shown to be suitably approximated by the MM environment on the one hand (4OT), and to be required to be included in the QM subsytem on the other (TET2). Additionally, various residues located in the 2nd– and 3rd– shell are conserved and are important for catalysis (Figure 3.4).

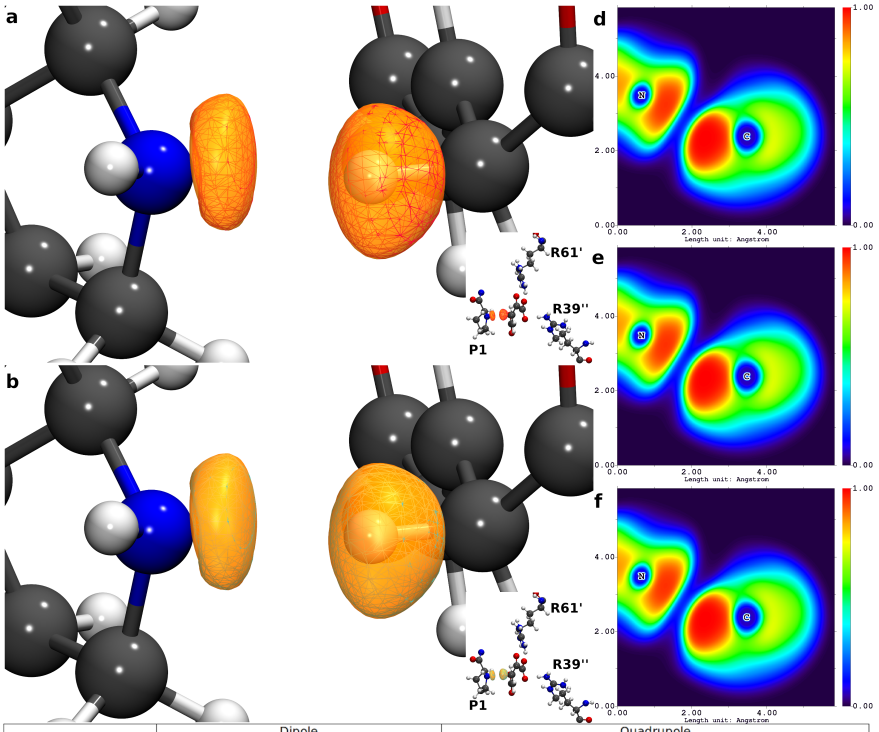
TopMod also provides the ability to generate modified wavefunction files for a subset of the atoms. We have performed ELF analyses for the reactant and TS structures of 4OT with a minimal active region and including R61' for the full wavefunction (see below, and Figure B.2), and a modified wavefunction including only the three atoms involved in the bond breaking/forming process. In all cases the ELF analyses result in nearly identical results. Therefore, for all other systems only reduced wavefunctions have been considered for computational efficiency.

3.3.3 Reactant Structure Analysis

As noted above, the structure of 4OT is composed of six monomers arranged in a trimer–of–dimers configuration that includes six active sites (Figure 3.1). 4OT exhibits high sequence and structure monomeric conservation (Figure 3.4). Each of the six active sites includes residues from three different monomers including the N–terminal catalytic P1, R11', R39" and R61'. The last two residues, R39" and R61', form hydrogen bond interactions (H–bonds) with the substrate. Additionally, L8 forms a backbone H–bond with the substrate that is important for catalysis.[139] All of these residues including L8, R11', R39" and R61' are conserved (Figure 3.4). Previous calculations showed that the minimum energy path (MEP) and associated free energy path (FEP) can be calculated with only P1 and the substrate in the QM subsystem.[60,61,99]

ELF calculations including multipolar decomposition (up to 2nd moments) based on wavefunctions from single point energies were performed for three different 4OT systems, including a minimal QM subsystem (P1 and substrate), one including P1, substrate and R39", and a third with P1, substrate and R61' (Figures 3.5 and B.1). In all cases, the total population for the basins associated with the atoms involved in the reaction is consistent between systems, with only minor changes ($\pm$0.02 e) for a small number of core and valence basins on heavy atoms. One exception occurs for the valence basin of O14 in the full wavefunction ELF analysis, which exhibits a difference of -0.1 for the R61' compared with the minimal systems.

The calculated first and second moments for the three 4OT systems are also similar, with maximum differences of $\pm$0.02 and $\pm$0.03 for individual components respectively. The basin corresponding to the bond that will be broken during the first step of the reaction (C2-H4) is largely unchanged between systems, with population remaining the same, dipole moment changing by -0.003, and quadrupole moment changing by -0.006 (Figure 3.5c). These data indicate that there is no appreciable difference on the electronic charge distribution for the reactant structure independent of whether the arginine sidechains are represented by the MM potential, or explicitly included in the QM subsystem.



c

	pop.	Dipole X.	Y.	Z.	Norm	Quadrupole XY.	XZ.	ZZ.	YZ.	X²Y²	Norm
MinQM	2.420	0.209	-1.480	1.056	1.830	0.136	-0.099	0.080	0.801	0.625	1.033
R39''	2.430	0.215	-1.514	1.055	1.858	0.162	-0.093	0.140	0.813	0.695	1.095
Diff.	0.010	0.006	-0.034	-0.001	0.028	0.026	0.006	0.060	0.012	0.070	0.062
R61'	2.420	0.204	-1.495	1.054	1.841	0.130	-0.084	0.102	0.807	0.652	1.054
Diff.	0.000	-0.005	-0.015	-0.002	0.011	-0.006	0.015	0.022	0.006	0.027	0.021

FIGURE 3.5. ELF analysis for 4OT reactant structures. **a)** ELF of reactive atoms comparing minimal QM (yellow surface) with R39" (red wireframe). **b)** ELF of reactive atoms comparing minimal QM (yellow surface) with R61' (cyan wireframe). **c)** Multipolar decomposition analyses for the valence basin of the transferring proton. **d)** 2D ELF Heatmap for Minimal QM. **e)** 2D ELF Heatmap for QM with R39". **f)** 2D ELF Heatmap for QM with R61'.

Three systems have been considered for TET2. The first involves a QM region comprised of the substrate and full first coordination sphere of the Fe: Fe(III), oxo (ferromagnetically–coupled, intermediate spin quintet), H1382, D1384, H1881, succinate, one 1st–shell H_2O, 5hmC, the 2nd–shell H_2O, T1372 and Y1902. The second system (no 2nd–shell H_2O) excludes the second shell water (Figures 3.2, B.2 and B.3), and the third system that was considered, excludes the conserved T1372 and Y1902 (no T1372/Y1902).

Figure 3.6 presents the ELF analysis for the three TET2 systems (see also Figures B.2 and B.3). Comparison of the system that includes the conserved T1372/Y1902 (large

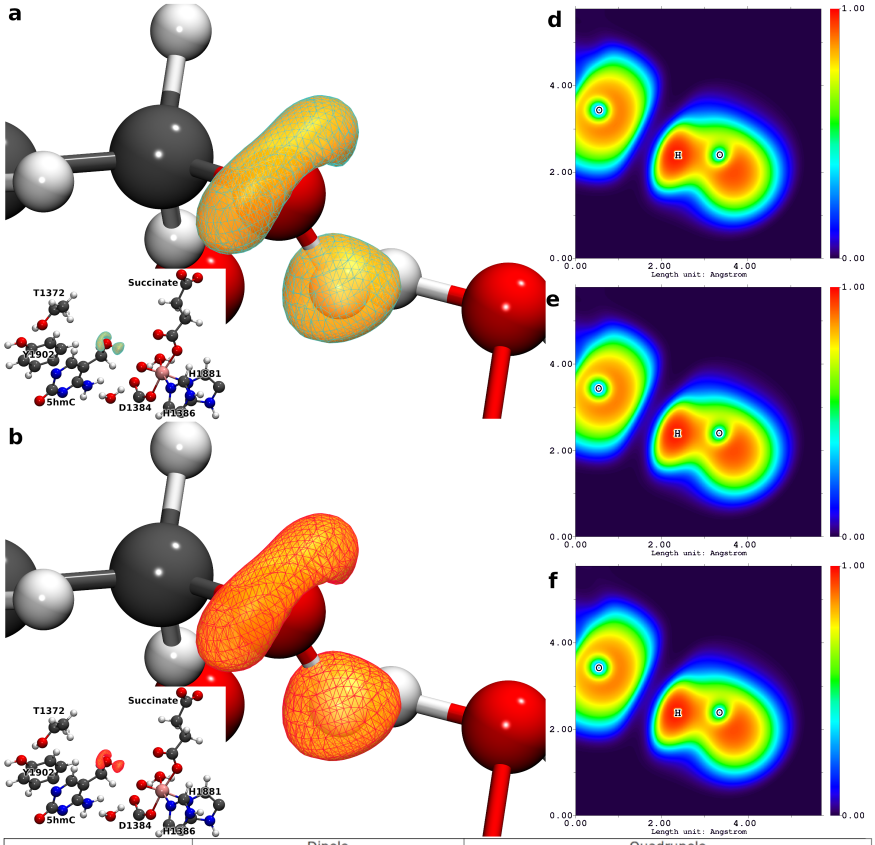
QM) with the system where these residues are represented by the MM environment suggests that these two conserved residues can be approximated by the MM environment without large changes on the electronic charge distribution on the reactive atoms. Comparing the TET2 active site with and without T1372 and Y1902 included in the QM region, we find that the population of the valence basin at the metal–coordinating oxyl changes by 0.04, the dipole moment is negligibly changed by 0.021, and the quadrupole moment is changed by 0.004 (see Figure 3.6c).

Conversely, the comparison of the ELF analysis for the TET2 system with all fragments in the QM region, with the TET2 system without the 2nd–shell water included in the QM region shows that the population of the basin at the metal-coordinating oxyl changes by -1.16, the dipole moment changes by a magnitude of -1.128, and the quadrupole moment changes by a magnitude of -1.367 (see Figure 3.6c). Interestingly, this change in the electronic charge distribution is not apparent from visual inspection of the 3D or 2D surfaces. Thus, the multipolar decomposition provides a more detailed insight on the electronic charge distributions on the ELF basins. These results indicate that the second–shell water has a significant effect on the electronic charge distribution. To investigate this further, we performed the same analysis for the TS structures (see below).

3.3.4 Transition State ELF Analysis

The ELF analysis for the first TS of the proton abstraction from 2o4hex catalyzed by 4OT is shown in Figure 3.7 (see also Figure B.4). The comparison of the ELF surfaces and multipolar decomposition show no differences in the electronic charge distribution between the minimal QM region compared with the systems that include either R39" or R61', similar to what is observed in the reactant. These results indicate that the classical potential provides a good approximation around the QM subsystem in this case. Moreover, these results are consistent with the experimental confirmation of the absence of an explicit general acid in the reaction, and instead, R39" only providing electrostatic stabilization for the 4OT catalyzed reaction.[120]

For the TET2 ELF analysis, similar effects to the reactant are observed. The TS

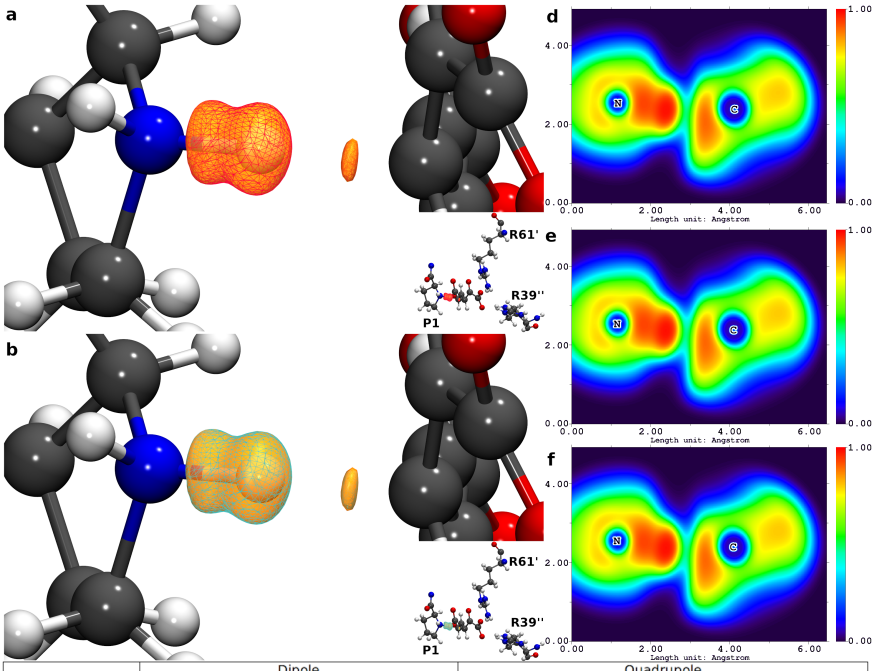


c

	Dipole					Quadrupole					
	pop.	X.	Y.	Z.	Norm	XY.	XZ.	ZZ.	YZ.	X²Y²	Norm
Large QM	3.900	-0.097	-2.975	-2.648	3.984	-1.245	1.011	-1.025	-4.150	2.225	5.078
No T1372/Y1902	3.940	-0.024	-3.005	-2.648	4.005	-1.045	1.142	-0.914	-4.062	2.468	5.082
Diff.	0.040	0.073	-0.030	0.000	0.021	0.200	0.131	0.111	0.088	0.243	0.004
No 2nd-shell H_2O	2.740	-0.706	-2.007	-1.906	2.856	-1.416	-0.647	-0.629	-2.876	1.637	3.711
Diff.	-1.160	-0.609	0.968	0.742	-1.128	-0.171	-1.658	0.396	1.274	-0.588	-1.367

FIGURE 3.6. ELF analysis of TET2 reactant structures. **a)** ELF of oxyl and O-H atoms on 5hmC (reactive atoms) comparing large QM (yellow surface) with QM excluding T1372/Y1902 (cyan wireframe). **b)** ELF of oxyl and O-H atoms on 5hmC (reactive atoms) comparing large QM (yellow surface) with QM excluding second shell water (red wireframe). **c)** Multipolar decomposition analyses for the valence basin of the oxyl atom. **d)** 2D ELF Heatmap for large QM. **e)** 2D ELF Heatmap for QM excluding T1372/Y1902. **f)** 2D ELF Heatmap for QM excluding second shell water.

structures for the 5hmC oxidation to 5fC catalyzed by wild type TET2 correspond to the hydrogen atom transferring from the hydroxyl O on the substrate to the oxyl atom on the ferryl.[62] The scaffold formed by the highly conserved T1372 and Y1902 residues (Figure



c

	Dipole				Quadrupole					
	pop.	X.	Y.	Z.	Norm	XY.	XZ.	ZZ.	YZ.	X^2Y^2	Norm
MinQM	0.700	-0.002	0.025	-0.023	0.034	0.018	-0.013	-0.028	-0.114	-0.110	0.162
R39''	0.700	-0.004	0.027	-0.026	0.038	0.025	-0.013	-0.027	-0.112	-0.112	0.163
Diff.	0.000	-0.002	0.002	-0.003	0.004	0.007	0.000	0.001	0.002	-0.002	0.001
R61'	0.700	-0.002	0.002	-0.003	0.038	0.007	0.000	0.001	0.002	-0.002	0.168
Diff.	0.000	0.000	-0.023	0.020	0.004	-0.011	0.013	0.029	0.116	0.108	0.006

FIGURE 3.7. **a)** ELF of reactive atoms comparing minimal QM (yellow surface) with R39" (red wireframe). **b)** ELF of reactive atoms comparing minimal QM (yellow surface) with R61' (cyan wireframe). **c)** Multipole Analysis for reactive proton. **d)** ELF Heatmap for Minimal QM. **e)** ELF Heatmap for QM with R39". **f)** ELF Heatmap for QM with R61'.

3.4) has been shown to be an integral part of the active site by orienting the substrate in the active site.[75] Comparing the ELF analyses of the TS structures when these two residues are included in the QM subsystem with the system where these residues are in the MM environment shows no changes on the basin associated with the transferring H atom as shown in Figure 3.8 (and Figure B.4).

On the other hand, significant differences are observed when looking at the system with the 2nd–shell water represented by the classical force field. In this case, a significant reduction in the population is observed with a concomitant change in the first and second moments of the basin. Here, the changes are such that the ELF basins for the lone pairs on

the hydroxyl O are significantly different (Figures 3.8b and B.5). Moreover, the representation of the water by the classical potential results in a significant overpolarization of the Fe atom inducing a subvalence splitting (see Figure B.5). This subvalence splitting effect on transition metals has been observed in other enzymatic systems.[140]

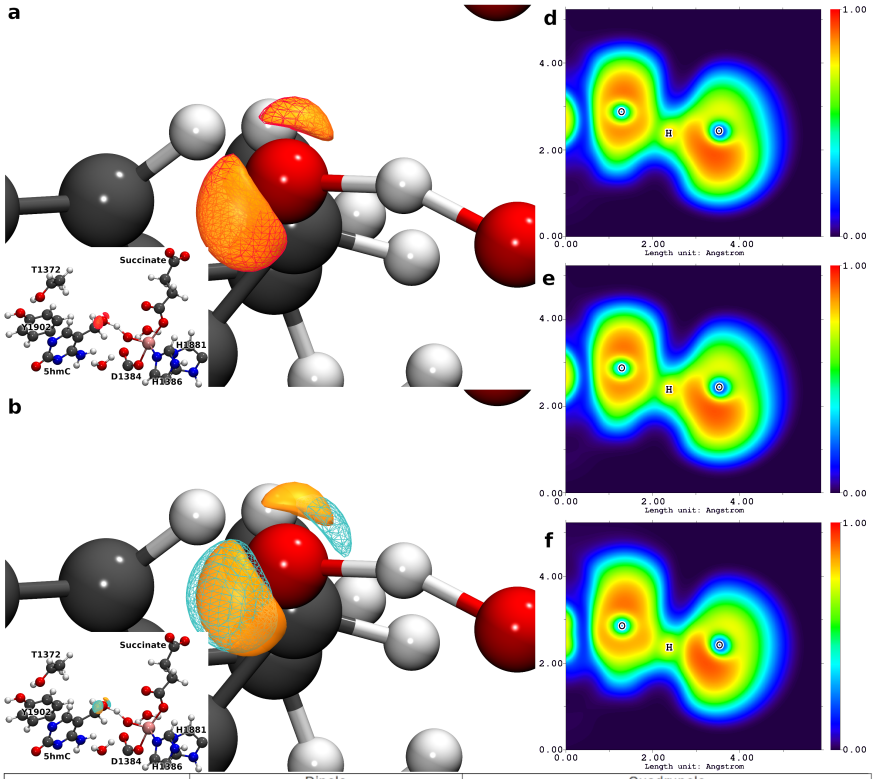
3.3.5 QM Subsystem Investigation for Polλ

To test the combined evolution/ELF approach, we have applied it to the investigation of the reaction mechanism of DNA synthesis by Polλ with three Mg^{2+} cations in the active site. Previous calculations that investigated the mechanism of Polλ involved only two metal cations.[63,140] These simulations indicated that several 2nd–shell residues are catalytically important, subsequently confirmed experimentally,[141] some of which are conserved across human polymerases (Figure 3.4).[63,100,127]

Additionally, the calculated reaction energy for the DNA synthesis step with two Mg^{2+} or two Mn^{2+} cations was reported to be endoergic.[63] As mentioned above, a third metal has been reported to provide stabilization for the product in two other polymerases, Polη and Polβ.[135–138] This stabilization for the product has also been observed for Polλ,[100,106] although the role of the third cation on the reaction mechanism of Polλ has not been investigated.

Therefore, we investigated the effect of three different conserved residues in the 2nd–shell of Polλ, R386, R420, and K472,(Figures 3.4, and S6) on the electronic distribution of the active site. Five systems have been tested including the original system (minimal QM region), three systems including only one of these 2nd–shell residues, and one more with all three residues included in the QM region.

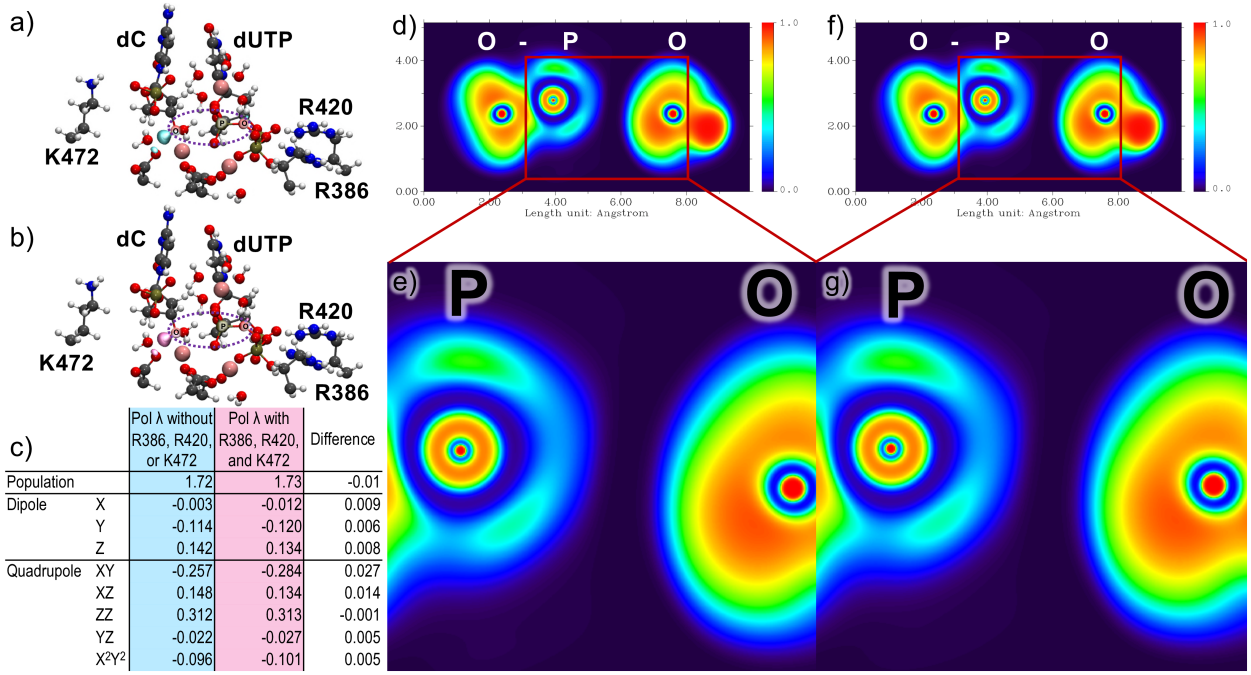
The ELF analysis for the original system, when compared with systems where a single one of these additional residues show no difference on the basins associated with the breaking of the Pα–O bond (Figure B.6). The most robust test of the evolutionarily conserved second–shell residues compared the original QM region against a QM region with R386, R420, and K472 included (Figure 3.9). The ELF analysis reveals minimal differences between the original system and the system including all three residues. The population



c	Dipole					Quadrupole					
	pop.	X.	Y.	Z.	Norm	XY.	XZ.	ZZ.	YZ.	X²Y²	Norm
Large QM	7.220	-4.049	1.378	-6.777	8.014	0.202	-6.897	-4.595	2.219	-1.302	8.678
No T1372/Y1902	7.180	-3.858	1.424	-6.785	7.934	0.385	-6.576	-4.801	2.216	-1.045	8.512
Diff.	-0.040	0.191	0.046	-0.008	-0.080	0.183	0.321	-0.206	-0.003	0.257	-0.166
No 2nd-shell H_2O	6.550	-6.947	-0.971	0.515	7.034	-2.892	1.098	3.451	-0.189	-6.538	8.016
Diff.	-0.670	-2.898	-2.349	7.292	-0.980	-3.094	7.995	8.046	-2.408	-5.236	-0.662

FIGURE 3.8. **a)** ELF of truncated wavefunction around reactive atoms in TET2 transition state with and without T1372 and Y1902 in the QM region. **b)** ELF of truncated wavefunction around reactive atoms in TET2 transition state with and without second shell water in the QM region. **c)** comparison of electron localization and multipolar moments of the iron-oxyl electronic basin for full QM region, QM without T1372 and Y1902, and QM without second shell water. **d)** ELF heatmap through region of reactive atoms Fe, O, H, and O with the full QM region. **e)** ELF heatmap through region of reactive atoms with T1372 and Y1902 represented as point charges in the MM region. **f)** ELF heatmap through region of reactive atoms with second shell water represented as point charges in the MM region.

changes by 0.01 in the large system for the reactive phosphate bond. The slight differences in the reacting bond's electronic environment engendered by the three residues either alone or collectively suggest that it is not necessary to include R386, R420, or K472 in the QM region.



c)

		Pol λ without R386, R420, or K472	Pol λ with R386, R420, and K472	Difference
Population		1.72	1.73	-0.01
Dipole	X	-0.003	-0.012	0.009
	Y	-0.114	-0.120	0.006
	Z	0.142	0.134	0.008
Quadrupole	XY	-0.257	-0.284	0.027
	XZ	0.148	0.134	0.014
	ZZ	0.312	0.313	-0.001
	YZ	-0.022	-0.027	0.005
	X^2Y^2	-0.096	-0.101	0.005

FIGURE 3.9. a) ELF for Polλ without R386, R420, and K472 in the QM region (blue). The reaction centers are outlined in purple. b) ELF for Polλ with R386, R420, and K472 in the QM region (pink). The reaction centers are outlined in purple. c) Comparison of electron localization with (pink) and without these three residues (blue). d) ELF heatmap of Polλ without additional amino acids on the plane passing through the reactive phosphate bond. e) Closeup of the reactive phosphate bond without additional amino acids in QM region. f) ELF heatmap of Polλ with R386, R420, and K472 on the plane passing through the reactive phosphate bond. g) Closeup of the reactive phosphate bond with three additional amino acids in QM region.

A path optimization of the DNA synthesis reaction with the third Mg^{2+} was performed using the quadratic string method (QSM) combined with a restrained MM proce-

dure as implemented in LICHEM.[107] The QSM path was approximated with 9 total images (7 structures between the optimized reactant and product) and optimized with an RMSD tolerance of 5×10^{-4} and 0.05 Å for the QM and MM subsystems respectively. The QM subsystem consists of 117 atoms including the original 2–cation system,[63] plus the new third cation, four water molecules, and 4 pseudobonds.[100,106]

The calculated minimum energy path including the third cation, without any of the 2nd–shell residues, is similar to the originally reported 2–cation mechanism. The MEP for the 3–cation system suggests a two–step mechanism with TS1 associated with the deprotonation of the O3' on the primer base, followed by the nucleophilic attack of the deprotonated O3' on the Pα of the incoming nucleotide. The calculated energy barriers for the approximate TS structures are 13.8 and 17.5 kcal/mol for the deprotonation and nucleophilic attack respectively (Figure 3.10. The rate–limiting barrier for the present mechanism is similar to the previously reported barrier of 17.6 kcal/mol with only 2 cations. These barriers are consistent with the experimental barrier approximated by transition state theory of 16.6 kcal/mol.[63] In addition, the calculated product (approximated from the optimized reactant) shows a difference of less than 0.2 Å RMSD compared with the experimental structure (see Figure Sx3), suggesting that the minimal QM subsystem is sufficient for this simulation.

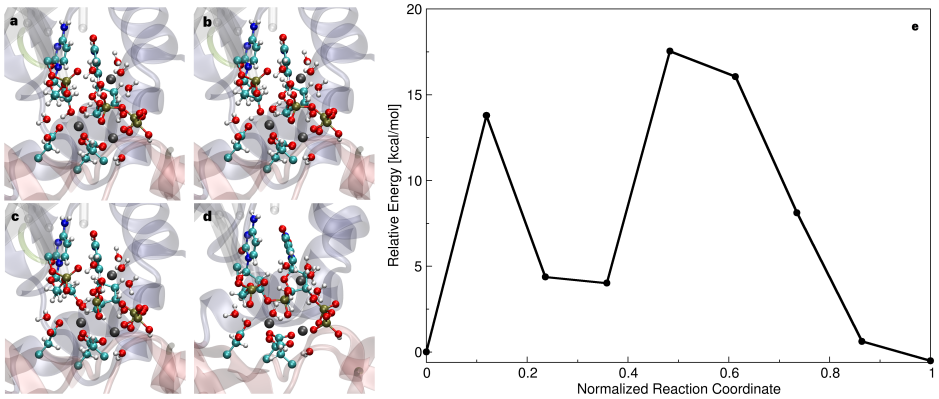


FIGURE 3.10. Critical structures along the MEP, a) reactant, b) TS1, c) TS2, d) product, and QSM optimized path e) for the DNA synthesis reaction catalyzed by Polλ with a third Mg^{2+} in the active site.

Two differences are observed between the previous and current reaction mechanism. The first is the fact that the calculated reaction energy is endoergic, ΔE=-0.5 kcal/mol, for the 3–cation mechanism compared with ΔE=2.3 kcal/mol for the 2–cation Mg^{2+} calculated path. Additionally, the present MEP indicates that the proton transfer is lower in energy than the rate–limiting step with an intermediate energy plateau in between the two barriers, whereas in the two–cation mechanism both steps were almost iso–energetic with a high energy plateau. Thus, the current results suggest that the thrid cation stabilizes the product, but does not affect the rate–limiting step in the DNA synthesis reaction catalyzed by Polλ.

3.4. Conclusions

A new procedure that combines protein sequence/structure evolution and ELF analyses has been developed to determine the size of the QM subsystem for QM/MM simulations. The use of evolutionary analysis provides a tool to determine potentially catalytically important protein residues. The use of ELF analysis provides a tool to investigate the effects of different environments on the electronic charge distribution of the QM subsystem, and how this distribution is polarized depending on whether a particular fragment is represented by the classical potential, or included in the QM region. Two systems, 4OT and TET2, were shown to provide positive and negative controls. In 4OT, a charged residue is appropriately approximated by the MM environment, even though it directly interacts with the substrate. For TET2, two highly conserved residues can be represented by the classical potential; however, a 2nd–shell water needs to be included in the QM region. The procedure was further tested by calculating the MEP associated with the DNA synthesis step catalyzed by Polλ with a third cation in the active site. The results indicate that three conserved residues around the active site can be represented by the MM environment, and the resulting MEP with the third cation is consistent with previous experimental results for Polλ and results for two other polymerases including a third cation.

CHAPTER 4

COMPUTATIONAL INVESTIGATION OF APOBEC3H SUBSTRATE ORIENTATION AND SELECTIVITY

4.1. Introduction

APOBEC3 enzymes (A3s) are cytidine deaminases responsible for dC→dU mutations. Most A3s act selectively on a 5'-dTdC-3' motif, with the exception of A3G which prefers a 5'-dCdC-3' motif.[142–144] A3H is known for its role in innate human immunity to retroviruses including HIV.[145] A3H acts upon a single-stranded DNA or RNA substrate and mutates cytidine to uracil, preferring a 5'-dTdCdA-3' sequence.[146] A3H is stable as a monomer in solution but is known to be an RNA-mediated dimer *in vivo*.[147] There are several available crystal structures of the A3H monomer, however none of the reported structures have been crystalized in complex with a substrate or substrate analog. Here, we present a computational investigation on the orientation of the substrate on A3H and possible structural determinants for the selectivity of the preferred 5'-dTdCdA-3' substrate motif.[146]

4.2. Methods

The initial crystal structure for monomeric A3H was obtained from the RCSB Protein Data Bank (pdbid: 5W45).[148] The sequence was confirmed to match the A3H-HapI consensus via Uniprot.[149] Protonation states of ionizable residues were assigned with H++, followed by system preparation with *tleap* in AmberTools16.[150,151] All systems were neutralized with Cl^- and solvated in water using a minimum distance of 12 Å between the protein surface and the edge of the box. The parameter sets employed were ff14SB,[152] OL15,[153] YIL,[154] and TIP3P.[155]

Molecular dynamics (MD) simulations were run in the NVT ensemble at 300K after minimization (50 steps with steepest descent followed by 450 steps with conjugate gradient)

This chapter is presented in its entirety from Hix, M. A.; Cisneros, G. A. Computational Investigation of APOBEC3H Substrate Orientation and Selectivity. *Journal of Physical Chemistry B* **2020**, *124* (19), 3903–3908. https://doi.org/10.1021/acs.jpcb.0c01857.

and iterative thermalization (20 equally spaced stages from 10K to 300K at 12,500 steps per stage) using a Berendsen thermostat.[156] Positional restraints with a force constant of 25.0 kcal mol^{-1} Å^{-2} were applied to the Zn^{2+}, coordinating residues, water in the active site, and the deoxycytidine, as the active site dissociated in unrestrained simulations. The cytidine base was held in the active site with a distance restraint of 15.0 kcal mol^{-1} Å^{-2}. Every system was simulated for 250 ns (in triplicate) with a 2.0 fs timestep and SHAKE for all bonds involving hydrogen atoms with the pmemd.cuda module in AMBER18 using a cutoff distance of 8.0 Å for nonbonded interactions and the smooth particle-mesh Ewald method for long–range Coulomb interactions.[157–159] Twelve systems (termed A–L) were generated by rotating the phostphate backbone of the ssDNA strand with respect to the central dC nucleotide in the active site (**Figure 4.1a**). The tested ssDNA sequence comprises 5'-dAdAdAdTdCdAdAdAdA-3'. All systems were built with the Modeller software.[160,161] Input coordinates and topologies included in Supplementary Information.

4.3. Results

Structural and dynamic properties, as well as average non-bonded (Coulomb and Van der Waals) residue–wise interactions via energy decomposition analysis (EDA) were calculated for each system (**Figure 4.1**) with cpptraj[162] and an in-house FORTRAN90 program (available in the ESI of Ref. **2**).[163–166] The EDA results suggest that system H has the most favorable total non-bonded interaction energy (**Figure 4.1b**). The ssDNA backbone for system B is oriented in the opposite 5'-3' direction to system H and has a lower interaction energy than adjacent systems. Root mean squared deviation (RMSD) analysis suggests that systems B and H have the smallest deviations with respect to the original structure, suggesting a more stable initial orientation of the ssDNA (**Figure 4.1c**, **Figure S1**). This is further supported by an RMS fluctuation (RMSF) analysis focused on the individual nucleotides in the substrate (**Figure 4.1d**, **Figure S2**). The average RMSF of the substrate for each system indicates that systems B and H have the smallest average fluctuations for the ssDNA substrate (2.0 Å and 2.1 Å respectively). Additionally it was observed that the systems adjacent to orientations B and H tended to shift the alignment of the DNA strand

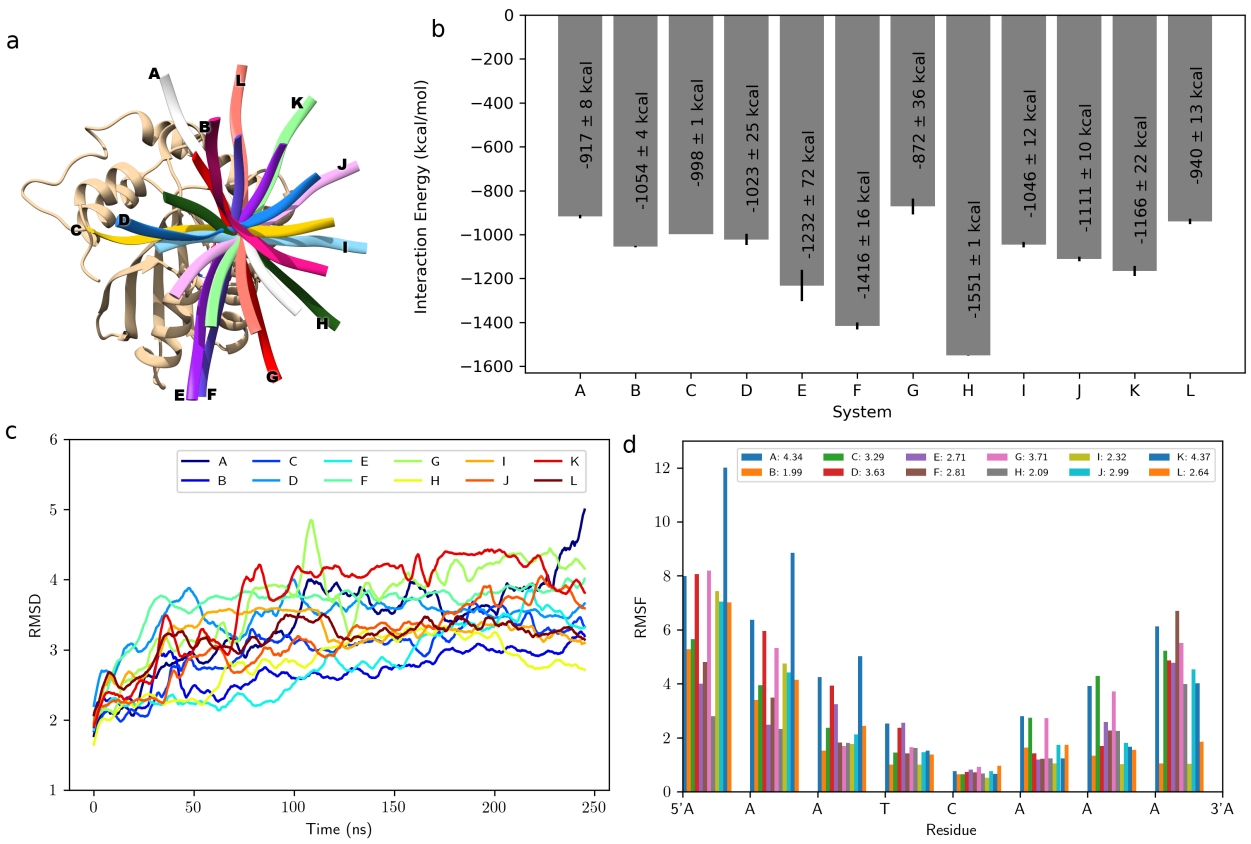


FIGURE 4.1. **a)** Twelve possible orientations of ssDNA substrate on A3H, **b)** Protein-substrate interaction energies with error bars showing standard deviation (calculated via the numpy python module using the first and last halves of the individual trajectories), **c)** RMSD of model substrate over time with respect to starting orientation, and **d)** RMSF of nucleotides in ssDNA substrate.

toward these two orientations. Analysis of residue–wise interactions with the entire ssDNA substrate suggests that systems B and H have the most favorable total interaction energies (**Figure 4.2**, **Figures S3 and S4**), with loop 1 ($_{17}$RRLRRPYYPRKALL$_{30}$) and the region comprising residues 115-130, which includes K117, K121 and R124, providing significant favorable interactions with the substrate. Taken together, these results suggest that the orientation of the phosphate backbone corresponding to the B or H systems allows for more favorable interactions with protein residues.

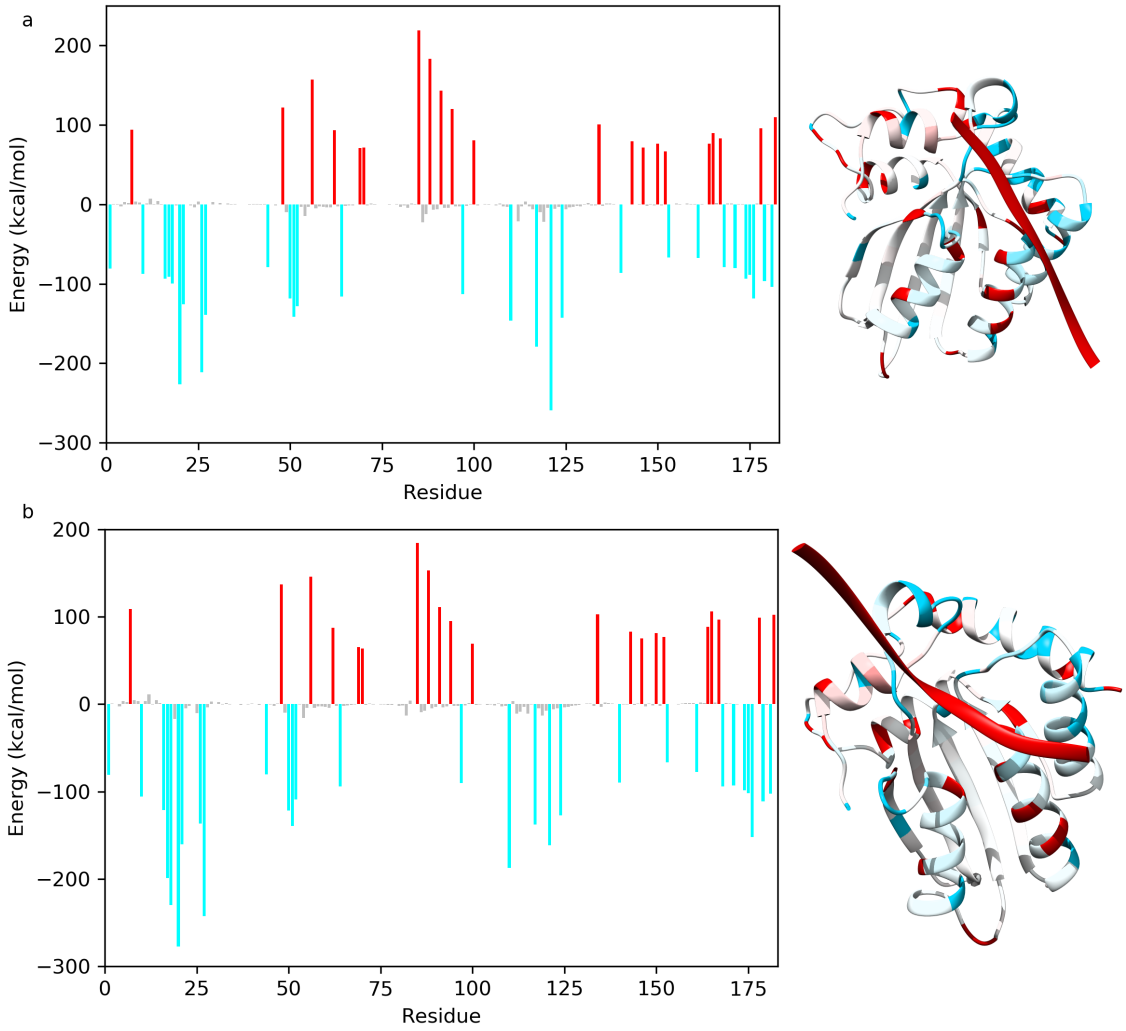


FIGURE 4.2. Non–bonded interaction energy between each residue and the ssDNA for a) System B and b) System H, with favorable(unfavorable) residue-substrate interactions in blue(red).

Systems B and H were run for an additional 250 ns from the end of the initial simulation with the restraints on the target dC removed to observe the stability of the enzyme-substrate complex. Interatomic distances between the reacting carbon on the cytidine and the water oxygen in the active site were measured over the trajectories to determine which system maintained a more stable binding. The ssDNA substrate in System B remained bound in the active site the entire duration of the simulation (average 3.2 Å). Conversely system H exhibited sporadic loss of binding. This suggests that while the total interaction

energy between the protein surface and the substrate is more favorable with system H, the local interactions with the target cytidine are more stable with system B.

Based on the above results, system B was selected for subsequent simulations to investigate the selectivity of A3H for the reported consenus sequence. Six new systems were generated by modifying the nucleotides that flank the central dC to investigate the effects of different bases in place of the preferred dTdCdA motif. Thymine recognition was tested by generating three system: 5'-dAdCdA-3', 5'-dCdCdA-3', and 5'-dGdCdA-3' substrates. Adenine recognition was tested by 5'-dTdCdC-3', 5'-dTdCdG-3', and 5'-dTdCdT-3' substrates.

EDA was performed to compare the differences (if any) in protein/substrate interaction between the 5'-dTdCdA-3' motif and all other systems (**Figure S5-S10**). The total non-bonded interaction between the protein and the nucleotides on the 5' position suggest that dT is favored over dG (-239 kcal mol^{-1}) and dA (-38 kcal mol^{-1}), but less than dC (+81 kcal mol^{-1}). The 5' dT nucleotide shows favorable interactions with R17, R21 and R26 when compared with those from purine nucleotides, however these interactions are generally unchanged when dC is in the 5' flanking position. dT also has more favorable interactions with S86 (between -1 and -5 kcal mol^{-1}) and less favorable interactions with S87 (between +1 and +5 kcal mol^{-1}) compared with the other nucleotides (see **Figure 4.3b**). The total non–bonded interaction of the protein with nucleotides in the 5' position shows that dA at this position is 5 kcal mol^{-1} less favorable than dT; dG is 61 kcal mol^{-1} less favorable, and dC is 68 kcal mol^{-1} less favorable. Hydrogen bond (HB) analysis indicates that a hydrogen bond is present between the 5' flanking nucleotide and S86 for 25.5 % of the simulation time in the case of a 5'dT, compared with 11.7 %, 0.8%, and 0.0 % for dA, dC, and dG respectively. The HB persistence between a 5' deoxy–nucleotide and S87 is 0.9 % for dT, compared with 0.2 % (dA), 10.7 % (dC), and 17.6 % (dG).

The total non–bonded interaction between the entire ssDNA strand and the protein is more favorable with dA in the 3' flanking position by between 31 and 245 kcal/mol compared with all other nucleotides at the same position. Our results suggest that 3'-dA shows strong attractive interactions with R17, R21, R26, K51, and K52. When compared with the other

nucleotides in the same position, the arginines interact more favorably with dA by at least 5 kcal mol^{-1} (see **Figure 4.3c**). These arginines are on Loop 1, which has previously been shown to be involved in DNA binding and recognition.[167–169] In contrast, the lysines show slight preference for all three other bases. A3H-R26 is structurally homologous to A3A-R28 and A3B-R211, which have been previously reported as key residues that drive selectivity in the respective A3s, and both preferentially act upon a 5'-dTdC-3' substrate.[143,170,171] The non–bonded interaction between the individual nucleotides at the 3' position and the entire protein suggest that dC and dT are less favored by 33 kcal mol^{-1} and 55 kcal mol^{-1} respectively, while dG shows $<$ 1 kcal mol^{-1} difference in interaction compared with dA. These results are consistent with previous experimental results showing similar selectivity for 5'-dTdCdA-3' and 5'-dTdCdG-3'.[146] Hydrogen bond analysis indicates an HB is formed between dA and R26 for 26.2 % of the simulation, compared with 0.8 %, 0.0 %, and 14.8 % for dC, dG, and dT respectively.

A3G, A3F and AID have a homologous loop, corresponding to residues 313 to 322 in A3G. Previous studies have shown that A3G$_{\mathrm{CTD}}$ 5'-dCdC-3' selectivity is driven by this loop and can be modified by mutating to the homologous AID or A3F sequences.[144,172,173] Based on our results, the DNA binding orientation precludes the interaction of the ssDNA substrate with the region in A3H that is homologous to the A3G 313–322 loop. One more difference observed between A3H and other A3s relates to the structure of the bound ssDNA substrate. In other A3s it has been reported that the DNA adopts a hairpin conformation.[143] In A3H, the substrate can orient in a hairpin conformation in a monomeric system, however, A3H forms an RNA–mediated dimer which obstructs this orientation.

Dimer simulations were carried out to test the possibility of the B, H and hairpin ssDNA orientations in the active site. The details of the system setup and simulations are reported in Ref **2**. The dimer structure shows that the active sites in each monomer are located near the RNA-mediated dimer interface of A3H. This location effectively prevents the ssDNA substrate from adopting a hairpin conformation. When a dimer system containing the superposed hairpin structure of A3A in both active sites was considered, the simulation

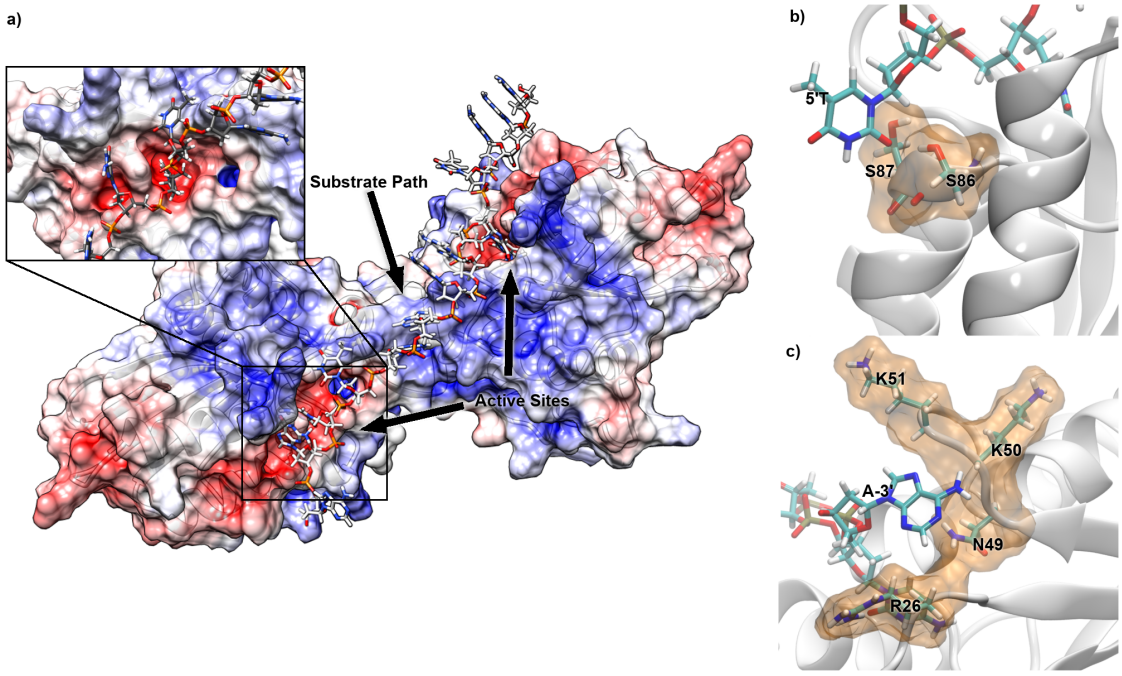


FIGURE 4.3. **a)** Electrostatic potential of A3H RNA-mediated dimer in solution with ssDNA mapped to solvent-accessible surface. Negative charges shown in red, positive charges shown in blue. ESP was calculated on the RNA-mediated dimer without model substrate using the APBS in PDB2PQR.[174] Inset is active site with substrate cytidine in position. **b)** Residues S86 and S87 interacting with the 5' thymine. **c)** 3' adenine in a pocket formed by R26, N49, K50, and K51,

was unstable due to strongly repulsive interactions between the ssDNA substrates and the RNA. Conversely, a track of positively charged surface residues is observed connecting the two active sites in the A3H dimer (**Figure 4.3**). For the B and H systems, the initial structures were simulated with two separate strands, one in each active site. During the initial stages of the simulations, both strands were observed to come together and thus a single strand spanning both active sites through the RNA interface was also considered as shown in **Figure 4.3a**.

4.4. Conclusions

In conclusion, we have performed computational simulations to investigate the binding orientation and substrate selectivity indicators for A3H. Our results suggest that the preferred binding orientation aligns the ssDNA substrate along a track that provides favorable interactions with the target cytidine and the two flanking residues, including three arginines that significantly favor the 3' flanking adenine. Our results are consistent with previous experimental reports on substrate binding and consensus sequence selectivity. Additionally some of the A3H residues predicted to be related to selectivity are homologous with selectivity residues reported for other A3s. These results also provide possible targets for mutagenesis to investigate the role of the selectivity filters for the consensus sequence of A3H.

CHAPTER 5

SINGLE-NUCLEOTIDE POLYMORPHISM OF THE DNA CYTOSINE DEAMINASE APOBEC3H HAPLOTYPE I LEADS TO ENZYME DESTABILIZATION AND CORRELATES WITH LUNG CANCER

5.1. Introduction

The eleven member Apolipoprotein B editing enzyme (APOBEC) family in humans are comprised of RNA and single-stranded (ss) DNA cytosine deaminases with diverse biological functions.[175–177] Within this family are a subgroup of seven APOBEC3 family members that primarily deaminate cytosine in ssDNA, which forms promutagenic uracil. This C-to-U conversion is considered a DNA lesion and can result in a C-to-T mutation if uracil is used as a template during DNA replication or error prone uracil repair can nucleate other mutations, such as C-to-G transversions.[178] These enzymes are known for their ability to restrict the replication of various viruses, such as HIV-1, Epstein Barr Virus, and Hepatitis B virus, through the mutagenic fates of uracil in DNA.[179–181] The host genomic DNA is usually safe from cytosine deamination due to various protections such as cytoplasmic localization, low expression levels, or cytoplasmic RNA binding which both sequesters and inhibits enzyme activity.[147,182–185] However, when these checks are not in place, the APOBEC3 enzymes that can enter the nucleus, such as APOBEC3B (A3B), APOBEC3A (A3A), and APOBEC3H Haplotype I (A3H Hap I),[81,146,147,186–189] can deaminate genomic DNA. This activity has linked these APOBEC3 enzymes to various cancers, such as lung, breast, bladder, cervical, and others. The activity related to cancer is on transient genomic ssDNA during replication or transcription and is considered "off-target" activity.[190–194] No-

This chapter is presented in its entirety from Hix, M. A.; Wong, L.; Flath, B.; Chelico, L.; Cisneros, G. A. Single-Nucleotide Polymorphism of the DNA Cytosine Deaminase APOBEC3H Haplotype I Leads to Enzyme Destabilization and Correlates with Lung Cancer. *NAR Cancer* **2020**, *2* (3), 1–27. https://doi.org/10.1093/narcan/zcaa023.

tably, the C-to-U conversion only occurs in a specific sequence context, e.g., 5' RTCA (A3B, "R" is A or G), 5' YTCA (A3A, "Y" is C or T), or 5'TCT (A3H).[146,188,189] This 5'TC sequence context has enabled the APOBEC3 mutation signature to be identified in at least 16 cancers and can be differentiated from chemical DNA damage, which may occur with lung cancer, for example.[178,195] The main hypothesis is that APOBEC3 enzymes create uracils randomly throughout the genome and that the majority of these are repaired by Base Excision Repair, but some may not be repaired, resulting in transition mutations, and some may be repaired in an error prone manner, resulting in transversion mutations.[196–198]

Since APOBEC-induced mutations are stochastic, the effects may be variable and can range from cell death, enhanced immune recognition, or tumor evolution.[146,178,188,197–200] These fates depend on the cell cycle and the location of the mutations. The interplay between APOBEC3 enzymes in this process is not known. Although there have been multiple reports of A3B, A3A, or A3H Hap I individually being involved in a cancer, it is not known if more than one APOBEC3 is expressed in a cancer cell at the same time and what would be the effects.[146,187–189,201] What has recently been identified is that APOBEC3 mutations occur episodically in cancers.[202] This is thought to be because constant expression and mutagenesis would result in cell death or recognition by the immune system, rather than creating a "just right" level of mutagenesis for cancer evolution. It has also been reported that for lung cancer, A3H Hap I predominantly causes early mutations and A3B predominantly causes late stage mutations.[146]

Interestingly, there are seven haplotypes for A3H, defined by polymorphisms or deletions at positions 15, 18, 105, 121, and 178, but only A3H Hap I is primarily localized to the nucleus.[203–205] The other A3H haplotypes are primarily in the cytoplasm.[205] A3H has several determinants for stability, such as the propensity to become ubiquitinated and ability to bind cellular RNA.[185,206] A3H forms a dimer that has no protein-protein contacts and is mediated by binding cellular RNA.[147,148,188,207] A3H haplotypes that stably bind RNA are located in the cytoplasm, and mutants that do not stably bind RNA are nuclear, suggesting that dimerization mediated by RNA suppresses potential off target deaminations by some

A3H haplotypes by maintaining cytoplasmic localization, since only those APOBEC3 members that have access to the nucleus can contribute to cancer mutagenesis.[147] In addition, some A3H haplotypes are rapidly ubiquitinated and degraded in cells (III, IV, VI), some are ubiquitinated and degraded on a slower timescale, enabling activity in cells (I), and others are not ubiquitinated and have a long half-life in cells (II, V, VII).[203–205,208] The short half-life of A3H Hap I is due to a G105, since mutation of this to an R105 forms an enzyme with a longer half-life in cells and is A3H Hap VII, which has been used as an A3H Hap I proxy in vitro.[175] How the amino acid at position 105 affects ubiquitination and cellular stability is not completely understood. Despite A3H Hap I being hypo-active in comparison to A3H Hap II, V, or VII,[208] it is the only A3H that has been implicated in cancer mutagenesis.[146] Starrett *et al.* have shown that A3H Hap I is an enzymatic contributor to 'APOBEC signature' mutations in lung cancer and likely contributor to breast cancer.[146] This was a surprising result since A3H Hap I has a half-life in mammalian cells of approximately 30 min.[203] The short A3H Hap I half-life also precludes the purification of significant quantities of A3H Hap I from eukaryotic cells for use in biochemical assays, although A3H tagged with Maltose Binding Protein fusion can be purified from E. coli and demonstrates activity.[208] When mammalian cell lysates are used and A3H Hap I is overexpressed at equivalent protein levels to more stable haplotypes, the activity is comparable.

One single nucleotide polymorphism (SNP), rs139298, resulting in A3H Hap I K121E has been recently reported to be associated with lung cancer.[209] Lung cancer results in over 1.6 million deaths per year and nearly 2 million new cases annually, with pulmonary adenocarcinoma comprising about 40% of the total cases of non-smoking lung cancer.[210–213] Surprisingly, we found that this K121E mutation further destabilized A3H Hap I, giving the counter-intuitive result that the absence of A3H Hap I is associated with lung cancer. In this study, we examine the A3H Hap I K121E variant and its stabilization by a novel A3H Hap I stabilizing mutation, K117E, using computational, biochemical, and cell-based techniques. Our results, considered in conjunction with the genetic data, strongly support the hypothesis that although APOBEC3-induced mutations can contribute to cancer evolution, constant

exposure to APOBEC3 mutagenesis may be detrimental to cancer cells.

5.2. Methods

5.2.1 Computational Methods

Molecular dynamics (MD) simulations were performed with the ff14SB,[152] OL15,[153] YIL,[154] and TIP3P[155] force fields using the pmemd.cuda[157,158] and OpenMM[214] programs. All systems were constructed from pdbid 5W45, corresponding to A3H-HapI based on the UniProt consensus sequence.[148,149] A search of 12 possible binding orientations of the DNA substrate on one single monomer was performed to determine the most probable orientation for the DNA substrate[1]

The most probable orientation was subsequently used to construct the biologically relevant dimer systems using the rhesus macaque-APOBEC3H dimer structure (PDBID: 5W3V) as a template for UCSF Chimera and taking into account the polarity of the substrate.[113,207,215] The monomer crystal of human A3H was overlaid on the dimer crystal of the macaque A3H and aligned. The protonation states for the dimer systems were determined using H++, neutralized with K^+ (Supplemental Table D.1) and solvated in a box of TIP3P water of approximately 143Å on each side to allow free movement without self-interaction across the periodic boundary.[150,151,216] Particle mesh Ewald was employed for long-range electrostatics, with a cutoff distance of 10Å and an error tolerance of 10^{-4} kJ mol^{-1} $Å^{-2}$.[158]

All simulations were performed in the NPT ensemble at 1.0 bar and 300 K using a Monte Carlo barostat, and a Langevin thermostat with a 1 fs timestep.[159,214,217] Simulation frames were saved at 10 ps intervals. To maintain active site geometry in the simulations, distance restraints of 20 kcal $mol^{-1}Å^{-2}$ at 2.0 Å were applied between the Zn^{2+} atoms and their respective coordinating histidines, with additional 10 kcal $mol^{-1}Å^{-2}$ restraints at 5.0 Å between the Zn^{2+} and water-coordinating glutamate to prevent incursion into the active site. Each system was minimized for 2,000 steps and equilibrated for 20 ns before beginning production. Additional simulations using distance constraints of 2.0 Å between the Zn^{2+} and coordinating atoms were also run and found to give quantitatively similar results to the

restrained simulations (data not shown).

Five systems were constructed including A3H Hap I wild type (WT), A3H Hap I K121E (cancer variant), A3H Hap I K117E/K121E, A3H Hap I K117E, and A3H Hap I G105R (A3H Hap VII). Every system was run for 250 ns in triplicate (750 ns total simulation time per variant). Energy decomposition analysis (EDA) was performed using an in-house FORTRAN90 program.[164–166] RMSF, RMSD, correlation, hydrogen bonding analysis, clustering, normal mode analysis, and distance calculations were performed using cpptraj in the AmberTools suite.[162,215]

5.2.2 Experimental Methods

For full experimental methods, refer to text of Ref. **2**

5.3. Results

5.3.1 Expression of A3H Hap I in a Lung Cancer Cell Line Induces γH2AX Foci

A3H Hap I has been implicated genetically in lung cancer, although no direct evidence of DNA damage has been shown in lung cells.[146] To better understand the A3H Hap I SNP (rs139298), we first characterized the level of DNA damage that could be induced by A3H Hap I WT. Since current models indicate that APOBEC3 enzymes do not induce cancers, but contribute to cancer mutagenesis in already transformed cells, we used NCI-H1563, a lung adenocarcinoma cell line derived from a male non-smoker.[146] We first determined whether NCI-H1563 cells expressed any endogenous A3s that have been implicated in inducing DNA damage. We found that NCI-H1563 did not express A3A, but did express A3B and very low levels of A3H mRNA, in comparison to the TBP mRNA control (Supplementary Figure D.2). The NCI-H1563 cells expressed 5-fold more A3B mRNA than TBP mRNA. The endogenous A3H was not genotyped since the expression was low. Instead, this provided an opportunity to transduce these cells with an A3H Hap I-Flag expression construct to make doxinducible stable cell lines to study the effect of A3H Hap I expression (Supplementary Figure D.3). A key marker of high A3 activity is replication fork stalling from base excision of uracil, leaving an abasic site, or formation of doublestranded (ds) DNA breaks that are induced by base excision of closely located uracils.[218] To avoid any possible cell death, we only

induced cells with dox for 24 h, and then cells were fixed and antibodies used to detect γ H2AX foci as a marker of stalled replication forks and dsDNA breaks.[219,220] The uninduced condition had low amounts of γ H2AX foci/cell suggesting that the A3B mRNA detected did not correlate with the protein level or the cell conditions inhibited high A3B activity (Figure 1A–D and Supplementary Figure D.2).[201] As a result, the cells were used to study the difference between uninduced and induced conditions that enabled expression of A3H Hap I-Flag (Supplementary Figure D.3). The uninduced A3H Hap I-Flag condition had 82% of cells having one to five γ H2AX foci/cell (Figure 5.1A–D). However, when A3H Hap I-Flag expression was induced with dox, the number of γ H2AX foci/cell increased with 12% of cells having over 15 γ H2AX foci/cell (Figure 5.1A–D, More, range was between 19 and 61 per cell). The other major populations were 24% and 58% of cells with 6–10 and 1–5 γ H2AX foci/cell, respectively (Figure 5.1D). The dox used to induce A3H Hap I-Flag expression was also tested in NCI-H1563 that were not transduced and showed that dox treatment by itself did not cause any γ H2AX foci to form above background (Figure 5.1A–D, Mock). DNA repair is responsible for inducing the stalled replication and dsDNA breaks after the formation of abasic sites from the BER enzyme, UNG, which removes the APOBEC-induced uracil. To confirm the connection of uracil formation to DNA damage, we conducted the experiment in the presence of a bacteriophage protein that is an inhibitor of UNG (UGI). In the presence of A3H Hap I-Flag and UGI, the γ H2AX foci/cell were similar to the uninduced condition, which demonstrates that A3H Hap I-Flag deamination activity forms uracils in DNA that are processed by DNA repair to induce abasic sites and/or DNA breaks that lead to γ H2AX accumulation. Interestingly, A3H Hap I-Flag did not induce any γ H2AX foci above the background in MRC-5 cells that are normal, fetal-derived cells (Supplementary Figure D.4). Although the reason for this is not known, we speculate that different DNA repair mechanisms in these cells may be responsible. These data demonstrate that A3H Hap I is able to induce formation of uracils that become DNA damage-inducing events after the action of UNG and, as a result, we sought to further characterize the effects of the A3H SNP resulting in the A3H K121E variant.

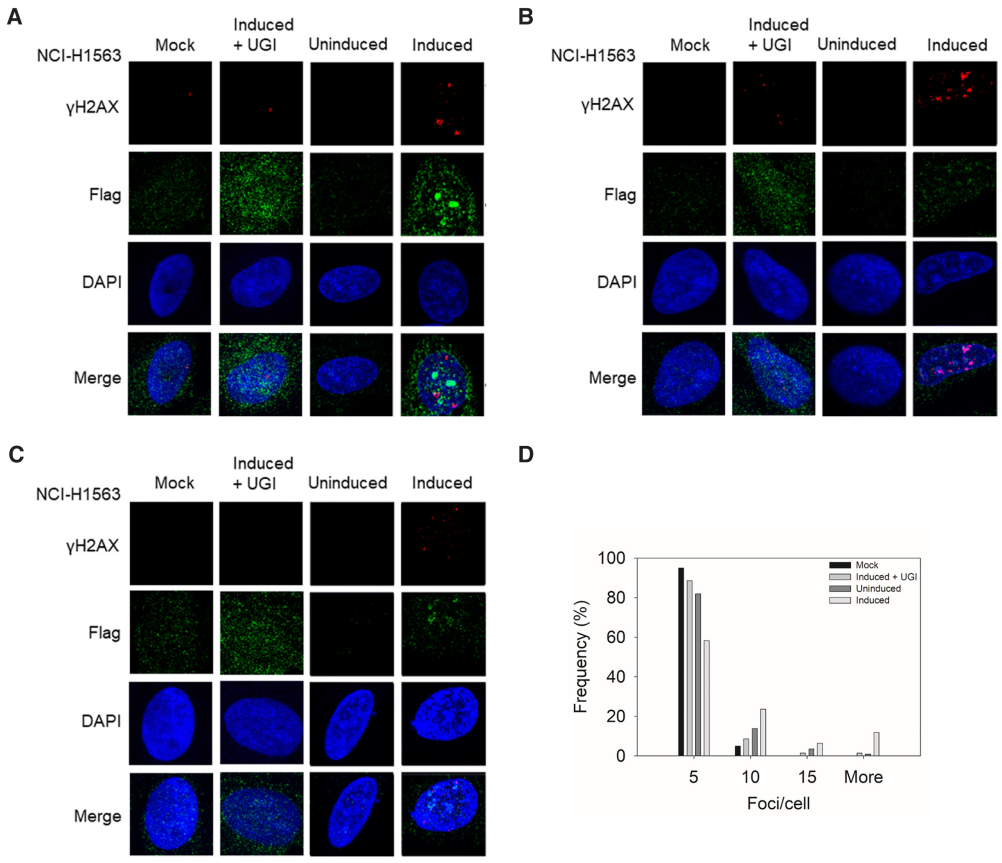


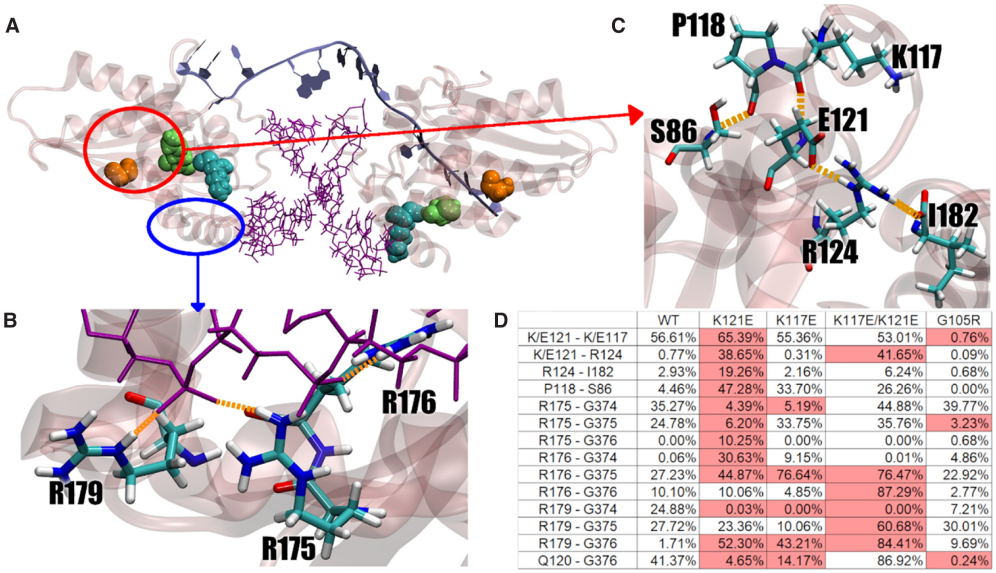
FIGURE 5.1. DNA damage induced by A3H Hap I. **(A–C)** The γ H2AX foci detected by immunofluorescence microscopy in NCI-H1563 cells that were or were not transduced to express a dox-inducible A3H Hap I-Flag protein. Different cell treatment conditions were dox in non-transduced cells (Mock), dox treatment to induce A3H Hap I, but with transfection of a plasmid express-ing UGI (Induced + UGI), transduced, but uninduced cells (Uninduced), and transduced and induced cells (Induced). The NCI-H1563 cells were analyzed 24 h after the treatment. Three representative images are shown. **(D)** Results were quantified and plotted as a histogram. Foci/cell are represented as 1–5 (Bin 5), 6–10 (Bin 10), 11–15 (Bin 15) and 15 (More) γ H2AX foci/cell.

5.3.2 A3H Hap I K121E Results in the Formation of a New Hydrogen-Bonding Network

To investigate the effect of the SNP resulting in the A3H Hap I K121E mutation on the protein structure and dynamics compared with A3H Hap I WT we used classical molecular

dynamics (MD). Pairwise hydrogen bond analysis of the K121E indicates the formation of a new hydrogen bond (HBond) network across part of the protein surface compared with the WT. This HBond network induces strain on the active site suggesting a change in stability or activity (Figure 5.2). Hydrogen bond interactions between protein residues that differed by more than 30% of the simulation time were identified (Supplemental Table D.2). The HBond network in the K121E system connects two existing networks (R124 to I182 and K117 to P118 to S87) from the WT into a single larger network in the K121E variant. This extended over-stabilized network effectively "freezes" the helix where the mutation occurs, disrupting both the active site and RNA-binding residues on the C-terminal helix. Changes in root mean squared fluctuation (RMSF) are also observed in the K121E system, with residues involved in the HBond network exhibiting reduced fluctuation while the rest of the protein exhibits increased fluctuation (Figure 5.2, Supplemental Figure D.5). Difference correlation matrices reveal that K121E has large regions of changing correlated movement, especially with respect to the region around the mutation point (Supplemental Figure D.6). Principle component analysis (PCA) shows that the first two modes of motion in the K121E system are constrained and tightly correlated in comparison to the WT (Figure 5.3, Supplemental Figure D.7).

The table in Figure 5.2d shows an increase in HBonds in several regions, including an increase in HBond prevalence between E121 and K117/R124. Based on this, we hypothesized that a second mutation that would eliminate the new interactions with E121 could rescue the cancer mutant. Thus, based on the increased HBond interactions from E121 to R124 and K117, these are the two likely candidates as mutation sites. In the case of R124 the HBond to E121 is formed via the sidechains. In addition, R124 also interacts with the C-terminus, I182. Thus, mutating R124 could result in negative effects. In the case of K117, the HBond to E121 is formed between the backbone atoms. Therefore, it was hypothesized that a mutant that would provide a repulsive electrostatic interaction between the sidechains at sites 117 and 121 could destabilize this backbone HBond interactions to levels similar to those observed in the WT, and restore the HBond network.



	WT	K121E	K117E	K117E/K121E	G105R
K/E121 - K/E117	56.61%	65.39%	55.36%	53.01%	0.76%
K/E121 - R124	0.77%	38.65%	0.31%	41.65%	0.09%
R124 - I182	2.93%	19.26%	2.16%	6.24%	0.68%
P118 - S86	4.46%	47.28%	33.70%	26.26%	0.00%
R175 - G374	35.27%	4.39%	5.19%	44.88%	39.77%
R175 - G375	24.78%	6.20%	33.75%	35.76%	3.23%
R175 - G376	0.00%	10.25%	0.00%	0.00%	0.68%
R176 - G374	0.06%	30.63%	9.15%	0.01%	4.86%
R176 - G375	27.23%	44.87%	76.64%	76.47%	22.92%
R176 - G376	10.10%	10.06%	4.85%	87.29%	2.77%
R179 - G374	24.88%	0.03%	0.00%	0.00%	7.21%
R179 - G375	27.72%	23.36%	10.06%	60.68%	30.01%
R179 - G376	1.71%	52.30%	43.21%	84.41%	9.69%
Q120 - G376	41.37%	4.65%	14.17%	86.92%	0.24%

FIGURE 5.2. Features of A3H Hap I WT and mutant enzymes. **(A)** A3H Hap I (transparent pink) with RNA interface (purple) and DNA substrate (ice blue). G105 is shown in orange, K117 in cyan and K121 in green, with corresponding residues on opposite monomer shown in transparency. **(B)** RNA-binding residues on terminal helix. **(C)** Formed HBond network as a result of K121E mutation. **(D)** Table of largest contributors in each HBond interaction, shown as percentage of total simulation time.

The simulation data (Supplemental Figure D.7-D.8) suggest that a K117E mutation could rescue the K121E variant by restoring the two hydrogen bonding networks. The K121E/K117E variant removes the HBond between the side chains of 121 and 117, and partially restores the remaining interactions in the network towards their WT levels (Figure 5.2). The K121E/K117E system shows less of a change in fluctuation, with residues in the network exhibiting near WT fluctuation, similarly to what is observed for the stabilized A3H Hap I G105R (A3H Hap VII) (Figure 5.3). The PCA shows that the first two modes are similar to those of the WT, as does the comparatively minor change in correlated movement.

5.3.3 K121E Destabilzes A3H Hap I

To test the computational results, we produced the A3H Hap I variants in 293T cells for cell-based analysis. To determine the steady-state expression levels of the A3H

Hap I WT, K121E cancer variant, putative rescue mutant K117E/K121E and rescue mutant control K117E, we transfected 3x HA-tagged A3H expression constructs into 293T cells. Immunoblotting showed that steady-state expression levels for the cancer variant were not detectable by immunoblotting, suggesting that the K121E mutation destabilized A3H Hap I WT (Figure 5.4A), in agreement with the computational predictions (Figure 5.2). In support of this, A3H mRNA levels measured by qPCR showed that A3H Hap I WT, K117E, K121E and K117E/K121E mRNA levels were not significantly different, demonstrating that the destabilization was at the level of protein (Figure 5.4A). We were also unable to detect any deamination activity for A3H Hap I K121E in cell lysates (data not shown). We found that the K117E mutation stabilized A3H Hap I in the presence of the otherwise destabilizing G105 amino acid (32) and, as predicted, can stabilize the cancer variant (K117E/K121E; Figure 5.4A). Since the destabilizing SNP, K121E, was associated with lung cancer,[209] these data suggested that A3H Hap I somatic mutations may be detrimental to cancer progression perhaps due to extensive mutagenesis leading to cell dysfunction or immune recognition. While beyond the scope of this study, the A3H Hap I K121E instability is consistent with A3H Hap I mutations being identified only early in cancer, the episodic nature of APOBEC3-induced mutations in cancer and the ability of A3H Hap I to cause DNA damage (Figure 5.1).[146,202,209]

5.3.4 A G105R Mutations Results in a More Active and Stable A3H than a K117E Mutation

To understand the reason for the instability of the A3H K121E variant, we conducted a biochemical analysis of A3H using an Sf9/baculovirus expression system. Although a K117E mutation can stabilize A3H Hap I, it is not a naturally occurring variant. The naturally occurring stabilization of A3H Hap I is a G105R mutation, which is known as A3H Hap VII.[203] The reason for the instability induced by G105 is not entirely understood. Computational models of the G105R system exhibit dynamic motion and HBond patterns similar to the WT, indicating that this mutation does not negatively affect the structural stability of the protein (see Supplementary Figures D.6–D.8 for G105R computational data). The G105R may simply increase the propensity of the protein to become polyubiquitinated

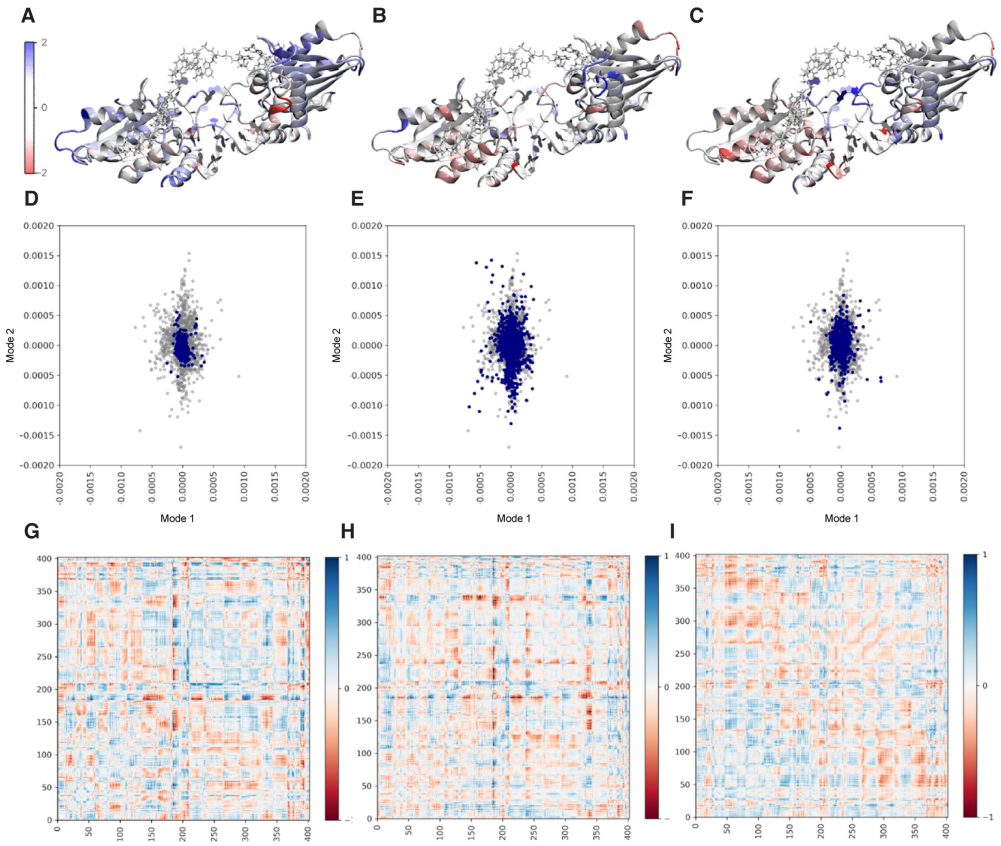


FIGURE 5.3. Difference RMSF for A3H Hap I WT and mutants. The analysis with respect to A3H Hap I WT overlaid on protein structures for **(A)** K121E, **(B)** K121E/K117E and **(C)** K117E. Residues in blue exhibit increasing RMSF compared to A3H Hap I WT; residues in red exhibit decreasing RMSF compared to A3H Hap I WT. PCA for mutant systems (blue) over WT (gray) for **(D)** K121E, **(E)** K121E/K117E and **(F)** K117E. Difference correlation matrices against WT reference by residue for **(G)** K121E, **(H)** K121E/K117E and **(I)** K117E. Correlation differences range from complete anticorrelation in red (-1) to complete correlation in blue (+1).

and degraded[206]), in contrast to K121E that destabilizes the structural integrity of A3H Hap I (Figures 5.2 and 5.3). To test whether the longer steady-state half-life of A3H K117E and K117E/K121E in cells correlates with increased catalytic activity similar to A3H Hap VII, we conducted an *in vitro* deamination assay using protein purified from Sf9 insect cells

to obtain a quantitative measurement. A3H deaminase activity was observed by adding a 118-nt ssDNA with two A3H deamination motifs (5' CTC) and an internal fluorescein label to purified A3H Hap II, A3H Hap VII, A3H Hap I K117E and A3H Hap I K117E/K121E. The two 5' CTC motifs account for the preference of some A3H variants to preferentially deaminate the cytosine motif nearest the 3' end of the ssDNA.[215] We found that despite recovery of protein stability the A3H Hap I K117E and A3H Hap I K117E/K121E were ∼4-fold less active than A3H Hap VII (relative to Hap I, G105R) or A3H Hap II (relative to Hap I, G105R, E178D) (Figure 5.4B and C, and Supplementary Figure D.9). These data suggest that the A3H Hap I HBond network is important for catalytic activity and compensatory mutations can correct the instability, but not fully reinstate catalytic activity.

A second component important for APOBEC3 activity is processivity, which describes the search process the enzyme undergoes to deaminate multiple cytosines in a single enzyme–substrate encounter. The low activity can be due to a less efficient search on the ssDNA, which precludes finding any cytosines for deamination or can be due to changes to the active site that influence the chemistry of the deamination.[175] The computational data predict that the active site structure would be affected, but cannot determine effects on activity or processivity, which are mitigated by ssDNA interactions with helix 6 on A3H.[215] The processivity assay is conducted under single-hit conditions, so only if an enzyme can undergo multiple deamination reactions in a single enzyme–substrate encounter, a double deamination band (5'C and 3'C) should be observed (see the 'Materials and Methods' section). The processivity factor represents the likelihood of a processive deamination occurring relative to a nonprocessive deamination, which for the A3H Hap I mutants is ∼6-fold. Consistent with the determinants of processivity being outside of the active site, the mutants were processive, similar to what has been reported for A3H Hap VII and A3H Hap II (Figure 5.4D).[215]

5.3.5 A3H Cancer Variant Destabilizes Dimer Interface

From monkeys to greater apes, the A3 family has lost activity. Although A3B activity can be lost through a gene deletion,[221] other A3s have lost activity because of decreased catalytic activity due to mutation of residues near the active site, e.g. A3D, or loss of

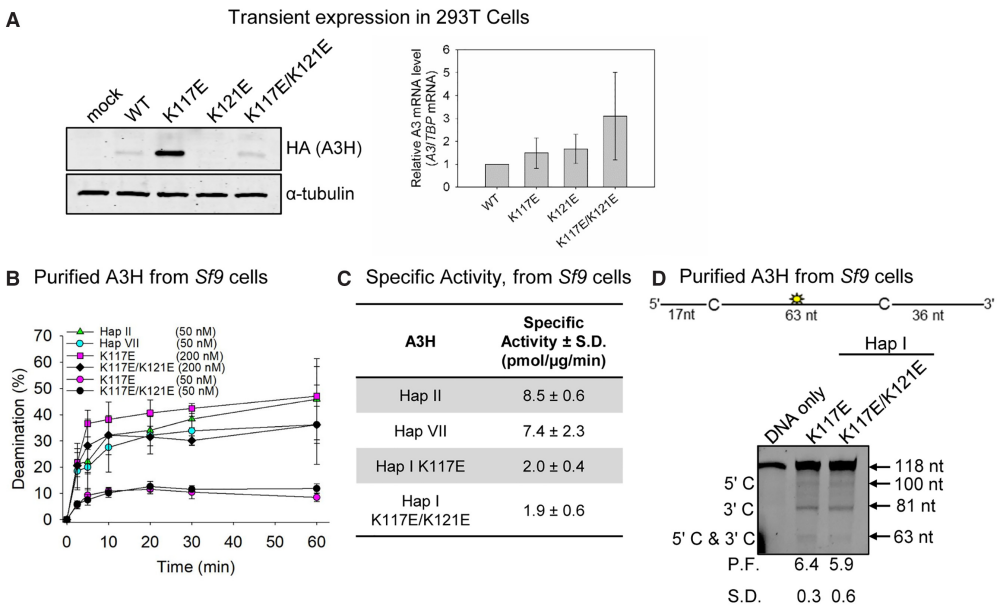


FIGURE 5.4. Expression and deamination activity of A3H and mutants. **(A)** Transient expression of HA-tagged A3H Hap I expression constructs in 293T cells was detected at the protein level by immunoblotting (left) and mRNA level by qPCR (right). The A3H Hap I WT, K117E, K121E and K117E/K121E had different steady-state protein expression levels, but the mRNA levels in cells were not significantly different, indicating that differences in stability were at the protein level. For immunoblotting, the α-tubulin served as the loading control. For qPCR, TBP mRNA served as the control and results are displayed relative to A3H Hap I WT. **(B)** Time course of deamination of A3H Hap II, A3H Hap VII and A3H Hap I mutants. Deamination was tested on a 118-nt ssDNA (100 nM) and gels analyzed for each plot are shown in Supplementary Figure S9. **(C)** Specific activity of A3H Hap II, A3H Hap VII and A3H Hap I mutants. **(D)** Processivity of A3H Hap I K117E and K117E/K121E. The processivity factor (P.F.) measures the likelihood of a processive deamination over a nonprocessive deamination and is described further in the 'Materials and Methods' section. Both K117E and K117E/K121E are ∼6-fold more likely to undergo processive deamination than a nonprocessive deamination. The standard deviation for three independent experiments is shown as error bars **(B)**, in the table **(C)** or below the gel **(D)**.

dimerization, e.g. A3C.[222,223] Based on recent crystal structures, the association of A3H with RNA is concomitant with enzyme stability and dimerization.[147,148,185,207,224] This is a unique feature of A3H compared to other A3s, in that it uses a dsRNA molecule to form a dimer interface without any protein–protein contacts and that this imparts stability to the enzyme. As a result, we hypothesized that the reason for the A3H cancer variant instability (Figure 5.4A) may be due to the inability to bind RNA and dimerize. RNA-mediated dimerization can be detected by treating a purified protein preparation with RNase A and observing whether an ∼12-nt RNA is protected from digestion. We confirmed that recombinant A3H Hap I mutants K117E and K117E/K121E purified from S*f*9 cells with RNA and RNase A treatment resulted in an ∼12-nt RNA being protected from degradation by the protein (Figure 5.5A). The A3H Hap VII and A3H Hap II also bound cellular RNA and protected an ∼12-nt RNA in the presence of RNase A, although the A3H Hap II bound less RNA than the other A3H variants. The reason for this is not known; although several studies have confirmed that A3H Hap II binds RNA in order to dimerize, the amount of bound RNA was not compared in parallel to other A3H haplotypes.[147,148,185,207,224] However, consistent with RNA binding, all four A3H variants formed primarily a dimer (44 kDa eluting at ∼17 ml, Figure 5.5B–D). Similar to A3G and as previously shown for A3H,[215,225] the size exclusion chromatography (SEC) profile peaks were broad, indicating polydispersity. There were smaller peaks on either side of the dimer peak, indicating smaller populations of monomers and tetramers. A3H Hap II and A3H Hap I K117E/K121E also had larger MW fractions than the other A3H variants (Figure 5.5D, 14–16 ml).

Since we cannot purify the K121E variant, we used computational analysis to predict whether the change at amino acid 121 would destabilize dimerization. RMSF analysis for WT, K121E and K117E/K121E (Figure 5.3) showed that residues in the K121E system exhibited increased fluctuation compared with WT, except those residues involved in the new hydrogen bonding network and many that interact with the RNA interface. These RNA-binding residues (R175, R176, R179) exhibit changing hydrogen bonding interactions with the RNA interface. In the simulations, these changes cause the RNA to change positions,

affecting additional protein-RNA interaction at loop 1 residues. Correlation matrices were calculated for all systems, with the K121E variant system showing increased correlation between the monomer subunits and decreased correlation within each monomer when compared to the WT (Supplemental Figure D.6), suggesting that the dimer interface is destabilized by the newly formed hydrogen bonding network. This prediction is supported by the SEC data which showed that the A3H Hap I K117E/K121E mutant had a distinct monomer peak, whereas A3H Hap I K117E had more polydisperse peak and was similar to A3H Hap I stabilized by G105R (A3H Hap VII) (Figure 5.5c-d). This was distinct from A3H Hap II that had the most predominant dimer peak (Figure 5.5c-d). This is in agreement with RMSF analysis where the K117E system shows qualitatively similar zones of correlation and anticorrelation at the dimer interface with respect to the WT. The K117E/K121E mutant correlation matrices reveal that this system more closely resembles the K121E than either the K117E or WT dimer systems, with qualitatively similar correlation matrices. These data are also in agreement with poor expression of A3H Hap I K121E in cells (Figure 5.4a). Further, consistent with the biochemical data, the K117E/K121E variant system exhibits fewer changes in fluctuation, and there is not a general trend of increased fluctuation as seen in the K121E variant.

5.3.6 A3H Variants Have Weakened Interactions with the Substrate

We observed that both the A3H Hap I K117E and K117E/K121E had ∼6-fold (K117E) and ∼2.5-fold (K117E/K121E) higher apparent dissociation constants from ssDNA than A3H Hap VII or A3H Hap II (Figure 5.5E–H). Further, both the A3H Hap I K117E and K117E/K121E binding curves best fit to a rectangular hyperbola by least-squares analysis, in contrast to A3H Hap VII and A3H Hap II that best fit to a sigmoid. The sigmoidal binding relationship for A3H Hap II has been shown to be due to further oligomerization of A3H dimers on ssDNA.[224] This does not appear to occur with A3H Hap I K117E or K117E/K121E (Figure 5E and F), although they can form dimers in solution (Figure 5.5B–D). The hindered oligomerization on ssDNA may be due to a faster off-rate from the substrate, exemplified in steady state by a higher apparent K_d (Figure 5.5E–H).

Altogether, the data indicate that the A3H Hap I K121E variant results in an inability to use RNA for dimerization, emphasizing the importance of cellular RNA to the structural stability of A3H. Further, the HBond network is important for interactions with ssDNA and oligomerization on ssDNA, providing a reasoning for the decreased catalytic activity in A3H Hap I K117E and K117E/K121E.

5.4. Discussion

A3H Hap I was identified to induce genomic mutations specifically in lung cancer.[146] Interestingly, the activity of A3H Hap I was found because researchers wanted to investigate how genomic mutations with a cytidine deaminase signature were still occurring in people with an A3B–/–deletion,[226–229] which occurs in the world population at 22.5%, although it is primarily found only in Oceanic populations.[221] Although A3A is also involved in cancer mutagenesis, it appears to be strongly associated with breast cancer.[201,230] The association of A3H Hap I expression and a mutation signature was strong in early clonal variants of lung cancers. Further supporting a role of A3H Hap I in lung cancer was a computational study that identified the association of SNP rs139298 in A3H Hap I with lung cancer.[209] In this work, we confirm a role of A3H Hap I in inducing DNA damage in lung cells (Figure 5.1). However, we also demonstrate in this work that the A3H Hap I cancer-associated SNP makes the resulting A3H Hap I K121E completely unstable (Figures 5.3 and 5.4A), suggesting that loss of this enzyme specifically promoted cancer.

While the effects arising from the combination of environmental, e.g. smoking, and A3 mutations remain unknown, the instability of the cancer-associated A3H Hap I K121E variant naturally leads to the hypothesis that if A3 enzymes cause too many mutations it would be detrimental to cancer evolution. Recent reports support the seemingly contradictory idea that A3 enzymes, specifically APOBEC3B, can both contribute to cancer and contribute to success of cancer immune therapy due to A3-mutation created neoepitopes[231] or efficiency of platinum-based drugs due to contributions to DNA damage.[232] This is usually also dependent on the type of cancer. In this report, we came to an equivalent conclusion since we determined that the previously identified rs139298 SNP that creates a K121E mutation

in A3H Hap I is associated with lung cancer and resulted in a loss of enzyme stability in cells (Figure 5.4A). While in many A3/cancer studies it is difficult to determine the fate of A3 mutations, the genetic data[209] combined with computational and biochemical study of the K121E mutation demonstrate that A3H Hap I has the potential for these detrimental effects on cell growth and proliferation in lung cancer.

The inactivation of A3 enzymes through SNPs is not unique to A3H, but A3H does have the highest number of inactivating SNPs in the A3 family. In particular, A3H has been lost multiple times in primate evolution, presumably due to nuclear localization and acquisition of genomic mutations.[203] While it is thought that A3H is kept active at a certain level in the population due to the antiviral effects of the enzyme, there are strong population stratifications in A3H active/hypo-active/non-active forms that correlate with the historical levels of HIV infection, a pathogen that active A3Hs can restrict efficiently.[233,234] The antiviral properties of these enzymes are likely why they are maintained despite the negative effect of their off-target activity. The destabilizing SNPs for A3H other than the one studied here are not thought to destabilize protein structure, but to promote A3H ubiquitination and proteosomal degradation.[206] Thus, the SNP studied in this work for A3H Hap I is unique since it appears to destabilize the enzyme by disrupting an HBond network. While this disruption decreased catalytic activity, it did not cause protein unfolding. Rather, it decreased the interaction strength with a dsRNA molecule that A3H acquires from the cell and uses to dimerize (Figure 5.3 and Supplementary Figure D.6).[147,148,185,207,224] This dimerization using an RNA intermediate is essential to A3H thermodynamic stability.[185,224] While the nature of dimerization for different A3s is unique, e.g. only A3H uses an RNA intermediate, an SNP that regulates activity through dimerization has also been found for A3C.[223,235] This type of functional inactivation enables rapid evolutionary toggling of activity for times when enzyme activity is needed again.[236]

The characterization of the A3H Hap I K121E variant and its stabilizing mutant, A3H Hap I K117E/K121E, demonstrates the importance of the A3H HBond network for catalytic activity and ability to bind RNA. Moreover, since the identification of distinct

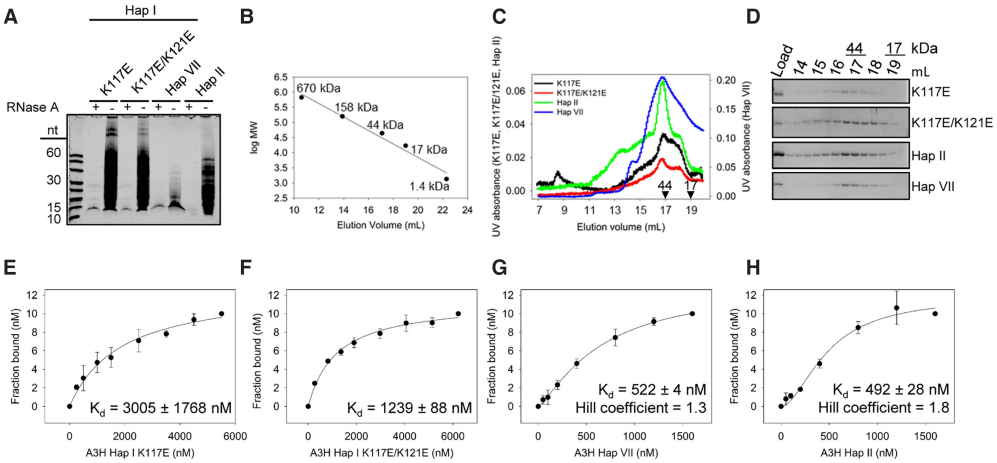


FIGURE 5.5. A3H Hap I mutant dimerization, oligomerization and ssDNA binding. **(A)** A3H dimerization is mediated by RNA. Purified A3H was or was not treated with RNase A and then denatured in formamide buffer and samples were resolved by urea denaturing PAGE. The gel was stained with SYBR Gold to detect nucleic acids. The A3H Hap I K117E, A3H Hap I K117E/K121E, A3H Hap VII and A3H Hap II protect an ∼12–15-nt RNA. Representative image shown from three independent experiments. **(B)** Standard curve and **(C, D)** SEC profile for A3H Hap I K117E, A3H Hap I K117E/K121E, A3H Hap VII and A3H Hap II demonstrating elution profiles that are composed of a dimer peak (44 kDa, 17 ml elution volume) and monomer peak (22 kDa, 19 ml elution volume). The 44 and 17 kDa elution volumes are shown on the **(C)** graph and **(D)** gels. **(E–H)** The apparent K_d of A3H enzymes from a 118-nt ssDNA was analyzed by steady-state rotational anisotropy. **(E)** The A3H Hap I K117E (3005 ± 1768 nM) and **(F)** A3H Hap I K117E/K121E (1239 ± 88 nM) bind the ssDNA with different affinities than **(G)** A3H Hap VII (522 ± 4 nM) and **(H)** A3H Hap II (492 ± 28 nM). The apparent K_d was calculated by determining the best fit by least-squares analysis, which was a hyperbolic fit for **(E, F)** and a sigmoidal fit for **(G, H)**. The Hill coefficient for A3H Hap VII was 1.3 and for A3H Hap II was 1.8. **E–H)** Error bars represent the standard deviation from three independent experiments.

mutational signatures and high A3 expression in cancers,[178,187,188] a number of research groups have contributed to our understanding of how A3 enzymes access ssDNA, which A3 enzymes are involved and, specifically for A3B, the clinical effects of the induced mutations.[189,190,192,194,237–239] The data presented here support the idea that A3H Hap I contributes to mutations in lung cancer. The combined genetic[209] and biochemical data presented here support the conclusion that the loss of A3H Hap I activity through the K121E variant may benefit the cancer and be detrimental to the host.

CHAPTER 6

DIVERGENCE IN DIMERIZATION AND ACTIVITY OF PRIMATE APOBEC3C

6.1. Introduction

Host restriction factors act as a cross-species transmission barrier for the Simian and Human Immunodeficiency viruses, SIV and HIV.[240] Overcoming these barriers by evolution of virally-encoded 'accessory proteins' which antagonize these restriction factors has characterized successful adaptation of the primate lentiviruses to new hosts.[241] In some cases, such as the infection of chimpanzees with SIV from Old World Monkeys (OWM) these adaptions to a new host was accompanied by major changes in the viral genome, which also facilitated the transmission to humans as HIV-1.[242] One of the major barriers to cross-species transmission of SIVs to a new host is the family of APOBEC3 (A3) host restriction factors that in most primates constitutes a minimum of seven enzymes, named A3A through A3H, excluding E.[243] The A3 family are single-stranded (ss) DNA cytidine deaminases that can inhibit retroelements, such as LINE-1, retroviruses, such as HIV and SIV, and some other viruses.[175,244–247] They belong to a larger family of APOBEC enzymes that have diverse roles in immunity and metabolism.[81]

For restricting HIV-1 replication, A3 enzymes first need to become packaged in the budding virion.[175,247] When these newly formed virions infect the next target cell, the packaged A3 enzymes deaminate cytidines on single-stranded regions of the (-) DNA to uridines during reverse transcription.[175,247] Uracil is promutagenic in DNA since it templates the addition of adenosine and results in G to A hypermutations on the (+)DNA when it is synthesized using a deaminated (-)DNA as a template.[175,247] This resulting double-stranded (ds) DNA provirus can become integrated but is non-functional.[175,247] Some A3 enzymes

---

This chapter is presented in its entirety from Gaba, A.; Hix, M. A.; Suhail, S.; Flath, B.; Boysan, B.; Williams, D. R.; Pelletier, T.; Emerman, M.; Morcos, F.; Cisneros, G. A.; Chelico, L. Divergence in Dimerization and Activity of Primate APOBEC3C. *Journal of Molecular Biology* **2021**, *433* (24), 167306. https://doi.org/10.1016/j.jmb.2021.167306.

can also physically inhibit HIV reverse transcriptase activity.[248,249] To counteract the actions of A3 enzymes, HIV encodes a protein Vif that becomes the substrate receptor of an E3 ubiquitin ligase with host proteins CBF-β, Cul5, EloB, EloC and Rbx2 to cause polyubiquitination and degradation of A3 enzymes via the 26S proteasome.[250] The members of human A3 family exhibit considerable variation in their antiviral activity; A3G can restrict HIV-1 ΔVif replication the most, with A3H, A3F, and A3D having decreasing abilities to restrict HIV-1 ΔVif.[203,235,251,252] A3A and A3B can potently restrict endogenous retroelements and some DNA viruses, respectively, but not HIV-1.[181,245,253]

A3C is the least active member of the human A3 family when it comes to restricting HIV-1 and HIV-2[254,255] because of its reduced ability to significantly deaminate (-)DNA due to a lack of processivity.[235] Since A3 enzymes deaminate only ssDNA and the availability of ssDNA during reverse transcription is transient, the enzymes must have an efficient way to search for cytidines in the correct deamination motif, e.g., 5'TTC for A3C, in order to maximize deaminations in the short time that the ssDNA is available.[175] However, a polymorphism in A3C that exists in about 10% of African individuals causes a change of S188 to I188, and this increases its HIV ΔVif restriction ability 5- to 10- fold, but it does not reach the level of restriction that A3G can achieve.[223,256] The increased restrictive activity of A3C S188I is due to acquiring the ability to form a homodimer, which enables processive deamination of HIV (-)DNA. Notably, the A3C S188I does not appear to be directly involved in the dimer interface, but instead induces conformational changes conducive to dimerization.[235] In addition, for the common form of human (h)A3C there is also a contribution of a specific loop near the active site of A3C (loop 1) to catalytic activity.[257] Interestingly, chimpanzee (c)A3C and gorilla (g)A3C despite having a S188 have equivalent HIV restriction activity to hA3C S188I.[235] This restriction activity of cA3C and gA3C is also attributed to their ability to form dimers, but through a different amino acid at position 115.[235]

In contrast to hominid A3C, the rhesus macaque (rh)A3C has been found to not be able to restrict replication of HIV-1, HIV-2, SIVmac (rhesus macaque), or SIVagm (African green monkey) in the majority of studies.[254,258,259] Paradoxically, the rhA3C contains

the I188 that enables activity of hA3C and enhances activity of cA3C, and gA3C.[223,235] Through direct coupling analysis (DCA)/coevolutionary analysis and subsequent Molecular Dynamics (MD), virological, and biochemical experiments, we uncovered a series of amino acid replacements required to promote dimerization in rhA3C and therefore promote HIV-1 restriction ability. Analysis of several OWM A3C sequences showed that they did not contain the "right" combination of amino acids for activity against HIV-1 suggesting that the evolutionary pressures that formed OWM A3C were different from hominid A3Cs that are active against lentiviruses. Overall, we form a model for the determinants of A3C antiviral activity and estimate its loss and gain throughout primate evolution.

6.2. Materials and Methods

6.2.1 Experimental Methods

For full experimental methods, see text of Ref. **5**.

6.2.2 Classical Molecular Dynamics

The initial structure of the human hA3C dimer was taken from the Protein Data Bank (PDB accession: 3VOW) and the peptide sequence was confirmed using Uniprot. The rhesus model (rhA3C) and all human-background variants were generated from the human dimer using Chimera to modify the peptide sequence.[113] Four single mutants (R44Q, R45H, N115M, A144S), a double mutant (R44Q/R45H), and two triple mutants (R44Q/R45H/N115M, R44Q/R45H/A144S) were all generated from the human wildtype (WT) background. MolProbity and H++ were used for all systems to determine protonation states of amino acids at pH of 7.0.[216,260–262] All models were neutralized to zero net charge with Cl^- and K^+ counterions and solvated using TIP3P water with a 12 Å minimum distance from protein surface to the edge of the solvent box.[155] This resulted in a solvated rectangular box unit cell measuring 82 Å x 84 Å x 108 Å with 90° angles between all adjacent edges. The AMBER FF14SB forcefield was used for the protein and TIP3P for water, Zn^{2+}, and counterions.[152,155]

Molecular dynamics simulations were performed using OpenMM on XSEDE's Comet HPC cluster.[214,263] Each system was equilibrated with iteratively reduced restraints on the protein to ensure stable simulation environment. The restraints began at 1000 kcal/mol

and were reduced by half after every completed stage of equilibration until the restraint fell to below 1 kcal/mol. Each completed stage was run for a minimum of 1.0 ns (106 timesteps) and checked for convergence to ensure stability of temperature (300K), density (1.0 g/mL), periodic box volume (approx 600 Å2), and total (potential and kinetic) energies of the system, resulting in a total equilibration time of 11.0 ns. To maintain active site geometry, harmonic restraints of 20 kcal mol^{-1} $Å^{-2}$ were applied between the active site zinc and the carboxylate carbon of E54 of each monomer subunit with an equilibrium distance of 5.0 Å. These restraints were maintained throughout the simulations. After equilibration, production dynamics were run for 100 ns using a 1 fs timestep and coordinates saved every 10 ps. A Langevin integrator was used as the thermostat with a Monte Carlo barostat in an NPT ensemble. The nonbonded cutoff distance was set to 10.0 Å, and the Ewald error tolerance was set to 10^{-3}. All models were simulated in triplicate for a total of 300 ns of production.

Analysis of MD trajectories was performed using cpptraj (correlated motion, normal mode analysis, root mean squared deviation (RMSD) and fluctuation (RMSF), hydrogen bond interaction analysis).[162] Data plots were generated with python, and 3D structure visualization was done using Chimera.[113] The first 100 normal modes were calculated and their relative contribution to the total motion was examined. Dimer interaction energies were calculated using the AMBER-EDA program[264] by calculating the ensemble average Coulomb and Van der Waals interactions between each amino acid on one monomer with each amino acid on the other.

6.3. Results

6.3.1 rhA3C is a Monomer, rather than a Dimer

Dimerization of hA3C, cA3C, and gA3C has previously been shown to correlate with HIV-1 restriction efficiency.[235] Since rhA3C has been reported as not being active against HIV,[254, 258, 259] we hypothesized that this could be due to a lack of dimerization. To determine the overall amino acid similarity, particularly at the dimer interface residues, the amino acid sequences for hA3C, cA3C, gA3C, and rhA3C were aligned (Figure 6.1A). The

rhA3C does contain an I188, which stabilizes dimerization in hA3C, cA3C, and gA3C (Figure 6.1A).[235] However, the other determinant identified for cA3C and gA3C that promotes activity against HIV-1 is K115. The rhA3C has a different amino acid at this position, M115, which is also different from hA3C that has an N115 (Figure 6.1A-B). Furthermore, amino acid R44 (and possibly R45) was predicted to interact with K115 in cA3C,[235] but in rhA3C these amino acids are Q44 and H45 (Figure 6.1A-B). The role of I188 in dimerization is indirect and the proposed mechanism for human A3C is that I188 causes a steric clash with F126 and N132 causing a repositioning of helix 6, enabling dimerization.[235] Due to the indirect nature of the role of I188 and that there are several key amino acid differences between hA3C and rhA3C, we purified the rhA3C wild type (WT) to determine its oligomerization state (Figure 6.1A-B).[235]

The rhA3C WT was purified from Sf9 cells in the presence of RNase A and we used size exclusion chromatography (SEC) to determine the oligomerization state. We found that the rhA3C WT had only one peak in the chromatogram that was a monomer (M, Figure 26.2-B), suggesting that the amino acid differences identified prevent dimerization. We then created rhA3C mutants to make it more hominid-like. We purified rhA3C mutants Q44R/H45R (hA3C-, cA3C-, and gA3C- like), Q44R/H45R/M115K (cA3C- and gA3C- like), and Q44R/H45R/M115N (hA3C-like). The rhA3C Q44R/H45R and Q44R/H45R/M115K mutants also only had one peak in the chromatogram that was a monomer (M, Figure 6.2A-B). However, the rhA3C Q44R/H45R/M115N had two peaks demonstrating formation of a dimer (D), but the monomer peak was more predominant (Figure 6.1A-B).

To determine if the rhA3C Q44R/H45R/M115N could also form dimers in cells we used co-immunoprecipitation (co-IP). Due to the propensity of A3s to bind RNA in cells and cell lysates,[265] the co-IP was carried out in the presence of RNaseA to ensure that interactions were due to protein-protein contacts and not mediated by an RNA bridge. Using two rhA3C constructs with either a 3xFlag- or 3xHA- tag we carried out a co-IP and found that the rhA3C Q44R/H45R/M115N was able to self-associate in cells. As a control, we also tested the single mutant M115N and this could not immunoprecipitate, suggesting that the

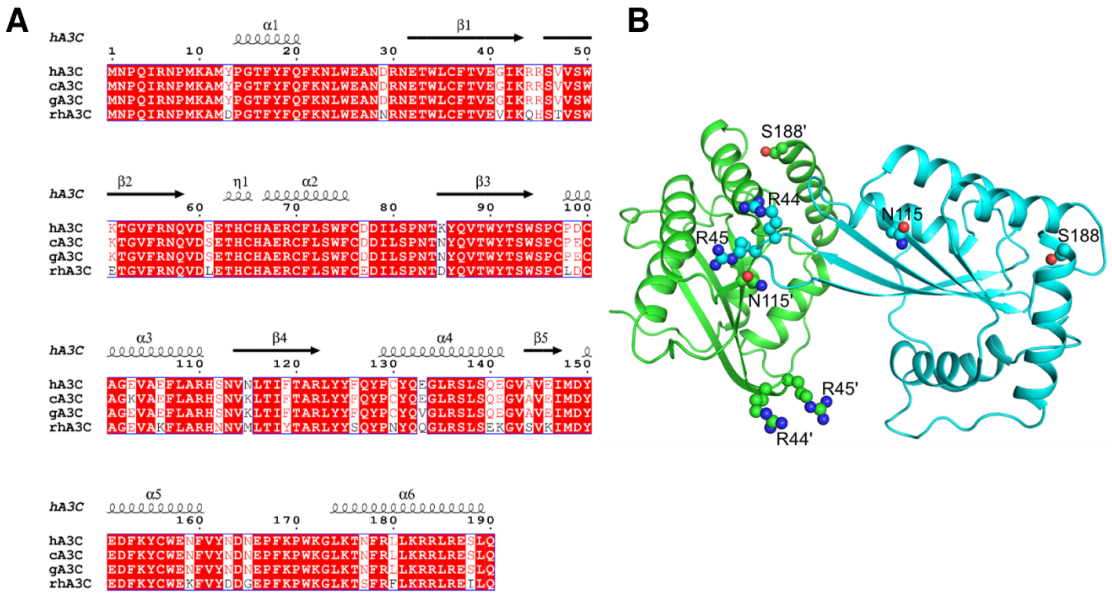


FIGURE 6.1. **Amino acid differences of rhA3C in comparison to hA3C, cA3C, and gA3C.** Sequence alignment and structural analysis of A3C. **(A)** Sequence alignment of hA3C, cA3C, gA3C, and rhA3C with amino acid differences shown in white. The sequence alignment was performed by a Clustal Omega multiple sequence alignment [50] and plotted using the program ESPript [76]. The α-helices and β-strands are shown above the alignment and are based on the hA3C 3VOW structure [35]. **(B)** The hA3C 3VOW structure is shown with amino acids important for hominid A3C dimerization shown on each monomer. The monomer shown in green has amino acids labeled with a prime symbol to differentiate them from amino acids belonging to the cyan monomer.

Q44R/H45R change enabled dimer formation (Figure 6.2C). Consistent with this, the rhA3C self-interaction was also found with Q44R/H45R and Q44R/H45R/M115K, but not M115K (Figure 6.2C). Thus, although in cells all three mutants with the Q44R/H45R change can form dimers, they do not all appear to be stable *in vitro* (Figure 6.2A-B). Collectively, the data suggested that amino acids 44 and 45 are key for rhA3C dimerization, but there may be different amino acids required to stabilize rhA3C dimerization than for hA3C.

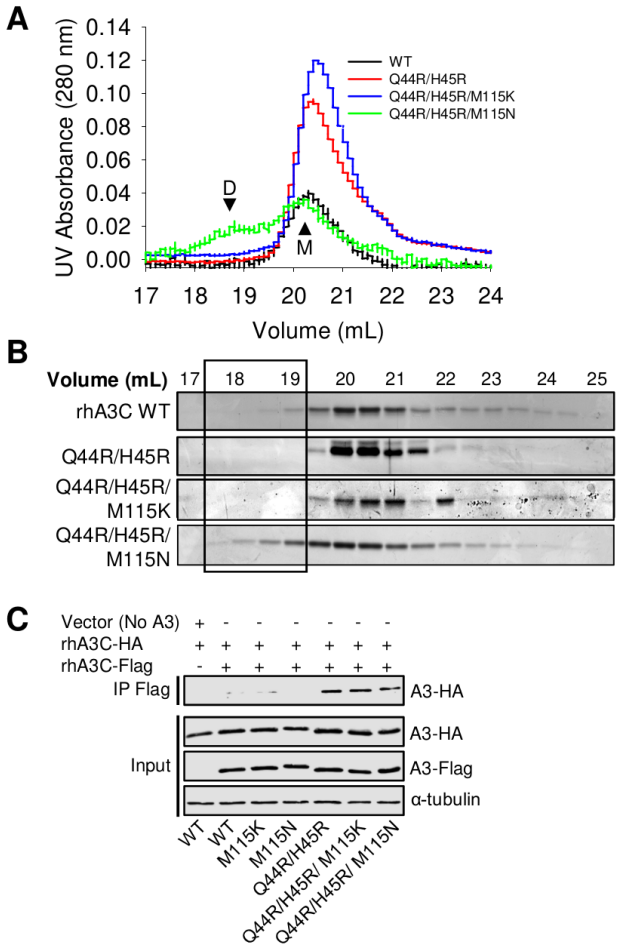


FIGURE 6.2. **Mutation of rhA3C amino acids to hA3C amino acids enables dimerization.**

6.3.2 rhA3C Deaminase Activity is Increased with Hominid-Like Amino Acids at Positions 44 and 45

Since the rhA3C appeared to have different determinants for stable dimerization than hA3C, we wanted to test that dimerization did indeed correlate with increased deamination activity in rhA3C. We examined the specific activity of rhA3C by conducting a uracil DNA glycosylase assay with a 118 nt substrate containing two 5'TTC motifs (Figure 6.3A-B). We used a time course of deamination to determine the linear range of activity (Figure 6.3A-B) and then calculated the specific activity of the rhA3C in that region (i.e., 5 to 15 min) (Figure 6.3C). We found that the rhA3C WT was approximately 2-fold less active than the three

rhA3C mutants containing the Q44R/H45R change, confirming that residues that promote dimerization are of primary importance to rhA3C activity (Figure 6.3C). However, since the rhA3C needs to search the DNA for the 5'TTC motifs among the non-motif containing DNA, the specific activity is made up of both the chemistry in the catalytic site once reaching the substrate C and the time it took to search for the 5'TTC motif.[175] The search occurs by facilitated diffusion and results in the enzyme being processive, i.e., deaminating more than one 5'TTC motif in a single enzyme substrate encounter. The processivity was calculated by analyzing the individual bands from the uracil DNA glycosylase assay (Figure 6.3B and 6.3D). The processivity requires single-hit conditions in which an enzyme would have acted only on one DNA (see Materials and Methods).

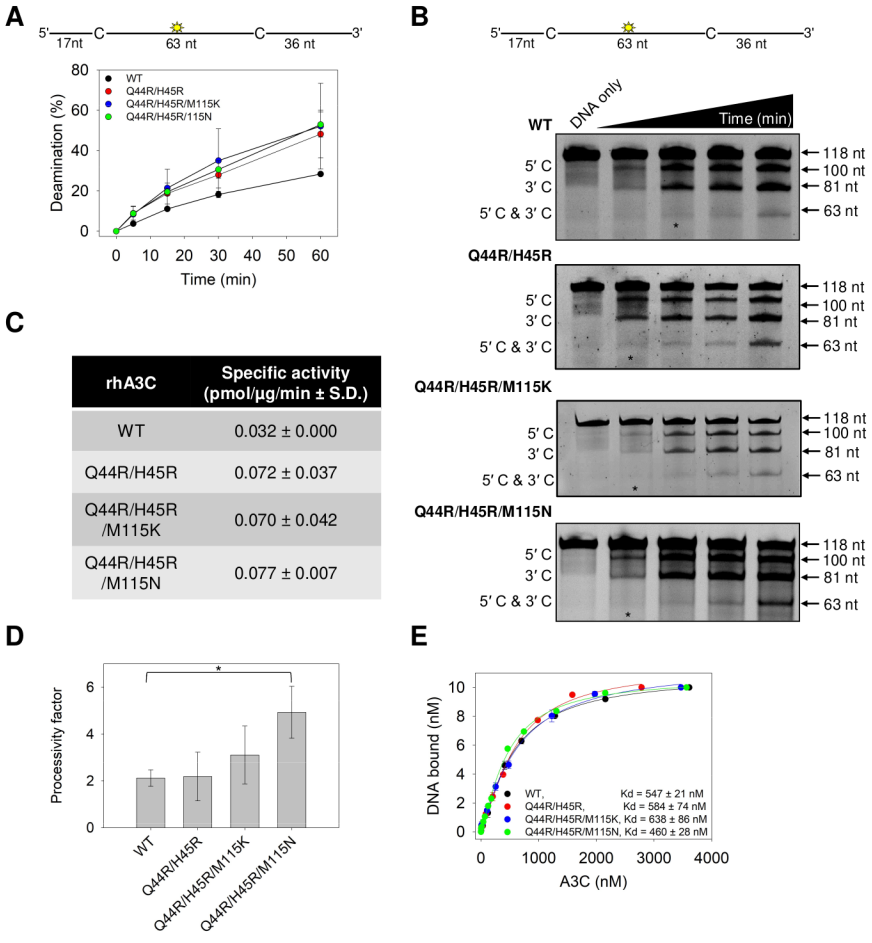


FIGURE 6.3. **Arginines at positions 44 and 45 in rhA3C increase enzyme activity.**

Using these conditions, we determined the number of deaminations at both 5'TTC motifs (5'C 3'C) compared to the single deaminations at either the 5'TTC motif closest to the 5'-end (5'C) or 3'-end (3'C) (Figure 6.3B). From these values a processivity factor can be calculated which is the fold likelihood that a processive deamination would take place. We found that the processivity of rhA3C WT was 2.1 (Figure 6.3D). In comparison, a nonprocessive enzyme has a processivity factor of 1. Thus, the rhA3C WT is not very processive, consistent with it being a monomer. The mutants had small but consistent increases in processivity where the Q44R/H45R processivity factor was 2.5, the Q44R/H45R/M115K slightly more processive (processivity factor of 3.2) and the Q44R/H45R/M115N that could partially dimerize most stably had a significant 2-fold higher processivity factor than the WT (processivity factor of 4.9, Figure 6.3D). However, dimerization did not increase the affinity of the A3C for the 118 nt ssDNA, as measured by steady-state rotational anisotropy (Figure 6.3E).

6.3.3 Increase in rhA3C Specific Activity Correlates with Increase in HIV-1 Restriction Activity

To determine if the *in vitro* differences of rhA3C from hA3C affect HIV-1 ΔVif ΔEnv (referred to as HIV) restriction ability we used a single-cycle infectivity assay. We compared hA3C S188I that is restrictive to HIV and rhA3C. We found that hA3C S188I could restrict HIV, but rhA3C could not, despite naturally having amino acid I188 (Figure 6.1A and Figure 6.4A). We also checked rhA3C activity against the OWM SIV from sooty mangabey monkey (smm). The rhA3C only restricted SIVsmm ΔVif 2-fold, in contrast to hA3C S188I (13-fold), hA3C (6-fold) and cA3C (4-fold) that were more restrictive (Figure E.2). Notably, the hA3C S188I was able to restrict SIV ΔVif better than HIV, suggesting that SIV is restricted more easily (Figure 6.4A and Figure E.2).[235] Thus, overall, the data show that rhA3C is not active against HIV or SIVsmm.

To determine if the lack of rhA3C restriction was due to it being a monomer, we tested the rhA3C mutants. We began by converting only the rhA3C Q44/H45 to R44/R45 (as in hA3C). We found that this mutation alone resulted in a 1.5-fold increase in restriction from

rhA3C wild type (WT) (Figure 6.4A). The WT and mutant were expressed and encapsidated into virions similarly (Figure 6.4B). Upon the addition of M115K and M115N mutations in a Q44R/H45R background, we found that there was a small increase in restriction. These triple mutants had a 1.7- to 2- fold increase in restriction activity against HIV in comparison to rhA3C WT, which was not due to increased encapsidation into virions (Figure 6.4B). Importantly, just the changes at amino acids 44 and 45 increased HIV restriction similarly to changes at position 115 (Figure 6.4A-B). Altogether, the data showed that the 2-fold increase in specific activity (Figure 6.3C) corresponded well with an approximate 2-fold increase in virus restriction (Figure 4A) and suggested that processivity may not be essential to HIV restriction by rhA3C. Further, these data demonstrated that rhA3C restriction activity was improved by converting dimer interface amino acids to the hA3C or cA3C amino acids, demonstrating that the rhA3C WT as a monomer is less able to restrict HIV. However, the restrictive activity of the triple mutants was still 2-fold lower than that of hA3C S188I (Figure 6.4A) which suggested that there were likely additional determinants for dimerization in rhA3C.

6.3.4 A Key Residue for rhA3C Dimerization is Uncovered via Direct Coupling Analysis

Since we could not achieve stable dimerization or HIV restriction activity equivalent to hA3C S188I from mutation of rhA3C amino acids 44/45/115 we turned to direct coupling analysis (DCA), also termed coevolutionary analysis.[266] Since dimerization is relevant for catalytic activity, we hypothesized that amino acid interactions that maintain function must be encoded in the evolutionary history of the A3 family of sequences (Pfam PF18771). The sequences included any organism and homologs that have a sequence that is classified as a member of the A3 family. A hidden Markov Model profile (HMMER) was used to search the NCBI database, including non-curated sequences, to find sequences whose statistics look like members of the family (Figure 6.5). This allowed inclusion of more distant sequences as opposed to direct matches to A3C. A resulting multiple sequence alignment (MSA) with 2500 sequences was compiled (see Materials and Methods and Supplementary File 1).

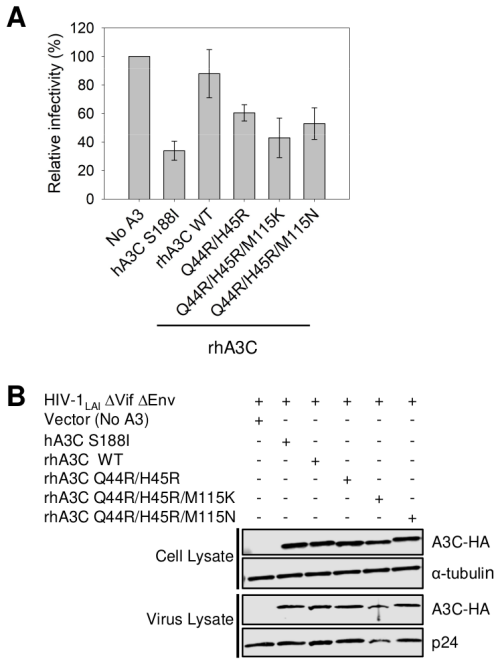


FIGURE 6.4. **Amino acids 44 and 45 in rhA3C are key determinants for HIV restriction ability.** **(A)** Infectivity was measured by β-galactosidase expression driven by the HIV-1 5 LTR from TZM-bl cells infected with VSV-G pseudotyped HIV ΔVif ΔEnv that was produced in the absence or presence of 3xHA tagged hA3C S188I, rhA3C WT, and rhA3C mutants Q44R/H45R, Q44R/H45R/M115K, and Q44R/H45R/M115N. Results normalized to the no A3 condition are shown with error bars representing the Standard Deviation of the mean calculated from three independent experiments. **(B)** Immunoblotting for the HA tag was used to detect A3C enzymes expressed in cells and encapsidated into HIV ΔVif ΔEnv pseudotyped virions. The cell lysate and virion loading controls were α-tubulin and p24, respectively.

Methods like coevolutionary analysis have been useful to identify amino acid coevolution in sequence alignments and to predict residue-residue contacts in the 3D structure of proteins[266,267] and to predict 3D folds.[268]

Interestingly, coupled amino acid changes that preserve dimeric interactions at the interface between oligomers can also be inferred and used to predict complex formation

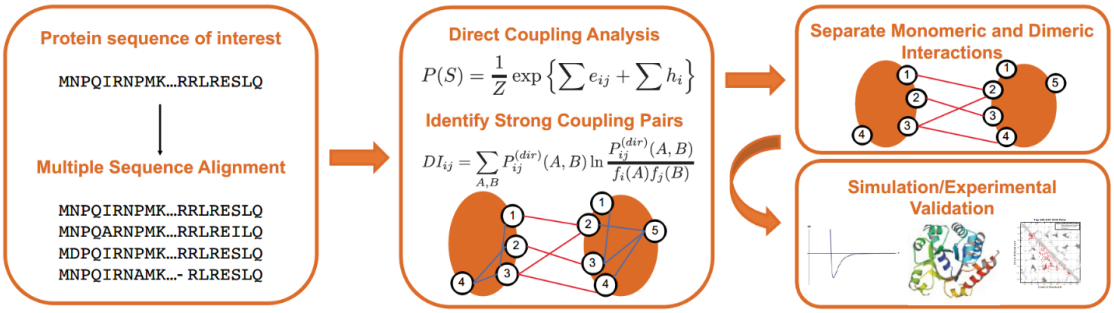


FIGURE 6.5. **Coevolutionary analysis to identify relevant interaction residues involved in dimerization.** First, a multiple sequence alignment (MSA) is created to identify members of the A3C family. The MSA is then processed using Direct Coupling Analysis (DCA) and a metric of coupling strength called Direct Information (DI). Once the top DI pairs have been identified, the monomeric crystal structure is used to distinguish monomeric from dimeric interactions. The resulting interacting residues are used to drive a simulation that predicts dimeric complexes. The residues involved in more coupled interactions are then proposed for experimental validation.

in dimers.[269–271] Therefore, we analyzed the alignment of approximately 2500 sequences (Figure 6.5 and Materials and Methods) to identify the most important dimeric coevolved interactions for the A3 family. Using DCA we identified a set of residue-residue pairs that are important for dimerization (Figure 6.6 and Materials and Methods). Figure 6.6A shows, in a contact map, the residue-residue contacts found in the crystal structure of hA3C (PDB 3VOW) for monomeric interactions (light gray) and the dimeric contacts (blue). In the same map, red dots indicate the top 300 coevolving interactions found by DCA. We notice that monomeric contacts can be predicted from the analysis of sequences, but more relevant to our study, we uncover a region (Figure 6.6B) that coincides with homodimeric contacts in the crystal structure (Figure 6.6C).

To further validate the relevance of those contacts for homodimeric interactions, we ran a coarse-grained MD simulation that uses such coupled coevolved pairs to drive a dimerization process (see Materials and Methods). The outcome of such simulation is a homodimer

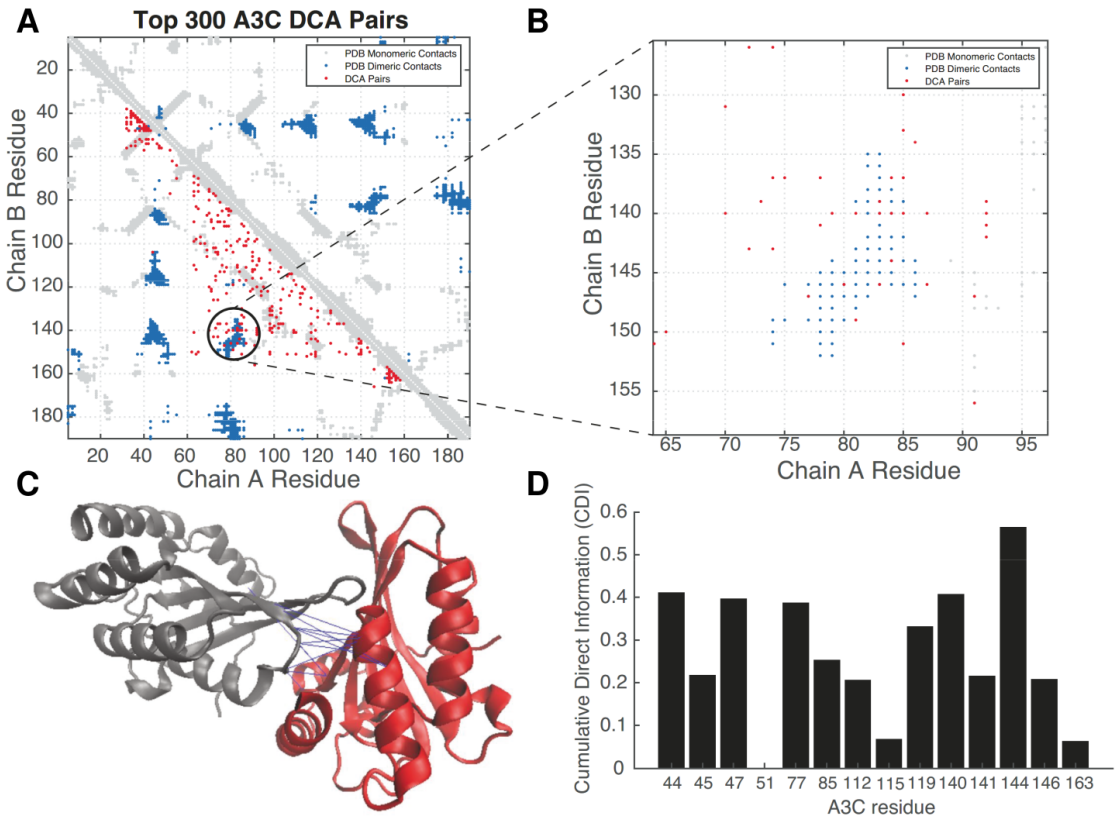


FIGURE 6.6. **Directly coupled residue pairs identify relevant dimeric interacting residues preserved through evolution.** **(A)** A contact map comparing the residue contacts in the x-ray crystal structure (3VOW) of hA3C (light gray for monomeric, dark blue for dimeric) and those interactions found by DCA (red). **(B)** A region of dimeric interactions shared by the crystal and coevolutionary analysis, highlighting the importance of those residues for dimer formation. **(C)** Overlay of the coevolved pairs on the 3VOW structure, depicting pairs inferred by the coevolutionary analysis. **(D)** A Cumulative Direct Information (CDI) metric identifies relevant residues at the interface. Of note, is the residue 144 that appears to have strong interactions with several residues and was therefore a candidate for further analysis.

predicted completely from coevolutionary signatures. We notice that this predicted complex deviates from the x-ray structure, but it does share a large portion of the homodimeric interface (Figure E.3).

Having identified a relevant coevolving interface for dimerization, we proposed a metric to identify key residues for dimer formation and that at the same time are part of the set of amino acid differences between hA3C and rhA3C. Our metric, called Cumulative Direct Information (CDI) quantifies how much a given residue is involved in coevolving interactions from the most important ranked residue pairs. We reasoned that if a residue is involved in several coevolving interactions, then they could be ideal candidates for mutational studies. This quantification for the top 200 dimeric interactions in the A3 family and for the residues that differ among the hA3C and rhA3C is shown in Figure 6.6D. We first noticed that the area between residues 44-47 was important according to this quantification, which was consistent with the experimental data (Figure 6.2). This analysis showed that residue 115 does not have a high CDI score, which agrees with our results of a limited impact of residue 115 on activity beyond residues 44 and 45 (Figure 6.3 and Figure 6.4). More importantly, Figure 6.6D shows a dominant role of residue 144 in dimerization and predicts residue 144 as a potentially important candidate for further mutational analysis.

6.3.5 MD Simulation Predicts an Important Role for Amino Acid 144 in Dimerization

To test the DCA and coarse-grained MD simulation we performed classical MD simulations using the dimeric hA3C crystal structure. MD simulations of hA3C tested the effect of converting the hA3C A144 to the rhA3C amino acid S144 in the context of changes at amino acids 44 and 45, forming the single amino acid hA3C mutants and the hA3C R44Q/ R45H/A144S triple mutant (rhA3C-like). If amino acid 144 is involved in dimeriza-tion then we would expect introduction of the rhA3C amino acid to disrupt the existing dimerization of hA3C. Consistent with this hypothesis, this mutant exhibited greater than 0.5 Å change in RMSF on 119 amino acids, and 9 changing by more than 1.0 Å with respect to the hA3C WT (Figure 6.7A). This corresponded to larger regions of change on the protein, most particularly on loop 1, which is located between α-helix 1 and β-strand 1 (Figure 6.1A and Figure 6.7A). Normal mode analysis based on the trajectories indicated that the essen-tial dynamics of all the systems are captured by the first two modes (Figures E.4 and E.5).

PCA revealed a clear change in the dynamic motion especially on the first mode (Figures E.6). These larger changes are observed alongside a noticeable reduction of average hydrogen bonding interactions at the dimer interface throughout the simulation (Figure 6.7B). The hydrogen bonds at the three mutation sites all showed considerable change, with position 144 increasing by more than 10% and positions 44 and 45 having six interactions decreasing by more than 10% (Figure 6.7B). The correlated motion of the variant is noticeably different compared with the hA3C, with a general trend of loss of correlated motion between the two monomers, consistent with a loss of dimer character (Figure 6.7C-D and Figure E.7). These changes in correlated motion are also consistent with the observed changes in RMSF (Figure E.8). Energy decomposition analysis indicates the dimer interaction is destabilized by approximately 770 kcal/mol (Table E.1). This destabilization is greater than the sum of the individual variants, consistent with what has been previously observed with multiple mutants.[272]

6.3.6 A rhA3C S144A Mutant in Combination with Q44R/H45R Enables Stable Dimerization and Increased Catalytic by a Unique Mechanism

Based on the predictions from coevolutionary analysis and MD simulations, we produced from Sf9 cells a rhA3C S144A mutant alone and in combination with changes at residues 44 and 45 to create a rhA3C Q44R/H45R/S144A mutant (hA3C-like). Consistent with the coevolutionary analysis, the S144A mutation enabled rhA3C to interact with itself (Figure 6.8A-B). However, a prominent dimer peak was only formed in the presence of a Q44R/H45R background (Figure 6.8A-B). Based on the hA3C crystal structure, we hypothesize that the rhA3C S144' pushes S46 away due to unfavorable non-bonded interactions (Figure 6.8C).[273] In contrast, the A144' mutant enables the S46 to come closer to the dimer interface, allowing the loop to properly orient and stabilize the dimer in conjunction with R44/R45 interactions (Figure 6.8C).

Consistent with the contribution to activity of amino acids 44 and 45 the rhA3C Q44R/H45R/S144A mutant, but not the rhA3C S144A mutant, increased activity 2-fold compared to rhA3C WT (Figure 6.9A-C). In addition, the rhA3C S144A, but not Q44R-

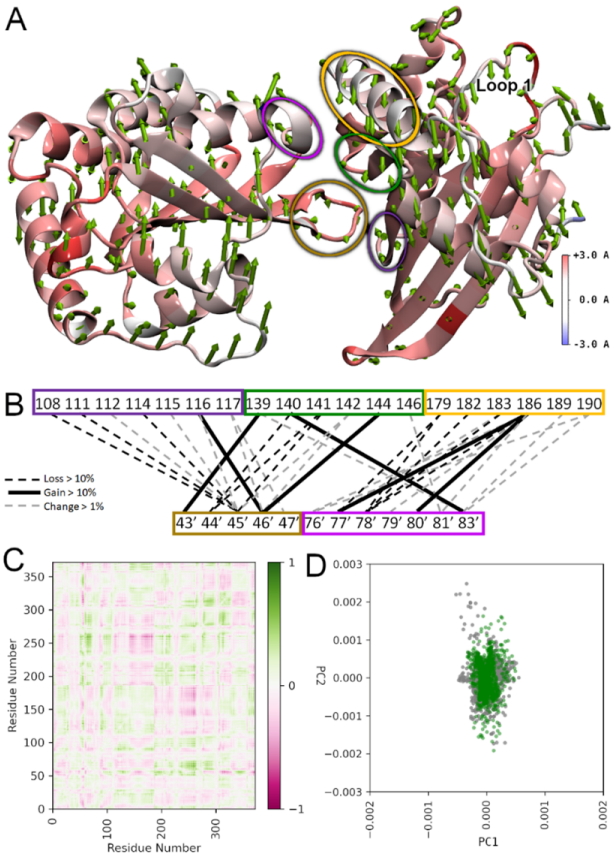


FIGURE 6.7. **Model analysis of hA3C R44Q/R45H/A144S shows large changes in specific conformations.**

/H45R/S144A mutant had a 1.5-fold increase in its K_d for ssDNA compared to rhA3C WT, suggesting that ssDNA binding is negatively affected when dimerization occurs in the absence of Q44R/H45R (Figure 6.9D). Surprisingly, for rhA3C Q44R/H45R/S144A, the increase in specific activity and maintenance of WT ssDNA binding affinity did not result in an increase in processivity (Figure 6.9E). Rather, the reason for this increased dimerization and catalytic activity without the expected increase in processivity appears to be due to changes in loop 1 (Figure 6.7). In multiple A3s, including A3C, loop 1 has been found to mediate accessibility of the ssDNA substrate to the active site.[257,274,275] The MD simulations revealed effects of changes at residue 144 on loop 1 that were unique from changes at residue 115 (Figure 6.7, Figure E.9, and Figure E.10). In the hA3C R44Q/R45H/A144S mutant, the increased RMSF on loop 1 corresponds to a reduction in dynamic motion compared with hA3C WT (Figure

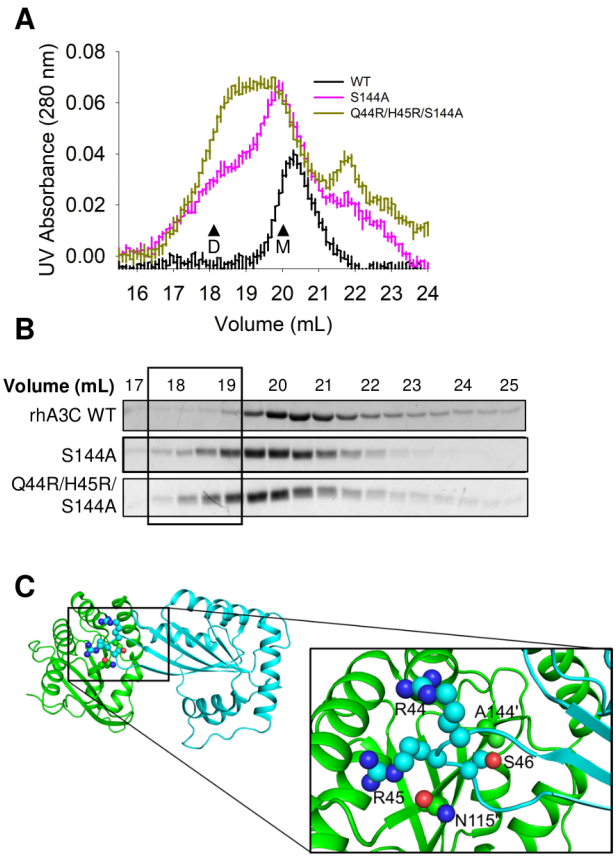


FIGURE 6.8. **Mutation of rhA3C amino acids 44, 45, and 144 to hA3C amino acids enables dimerization.** **(A)** SEC profile for rhA3C WT and rhA3C mutants S144A, and Q44R/H45R/S144A. Elution profiles for rhA3C WT was composed of a monomer peak (M, 20 ml elution volume). The rhA3C S144A showed a monomer and dimer peak (D, 19 mL elution volume). Only for the rhA3C Q44R/H45R/S144A there was a more prominent dimer peak. The elution profiles are shown as the UV absorbance during SEC elution. **(B)** Coomassie stained protein fractions resolved by SDS-PAGE that correspond to the eluted fractions in (A). Box shows where the dimeric fractions eluted. **(C)** The hA3C 3VOW structure is shown with amino acids important for rhA3C dimerization shown on each monomer. The monomer shown in green has amino acids labeled with a prime symbol to differentiate them from amino acids belonging to the cyan monomer.

6.7A and Figure E.9). The helices around the active site maintain similar RMSF across the hA3C WT and hA3C R44Q/R45H/A144S mutant consistent with a specific change in loop 1 mediating the changes in specific activity (Figure 6.7). Namely, loop 1 in hA3C is in an open/transition/closed conformation for 2%/80%/19% of the simulation time, while for R44Q/R45H/N115M this changes to 13%/23%/64%, and in R44Q/R45H/A144S this changes to 3%/36%/61%. In the hA3C WT and R44Q/R45H/N115M variant, there is little to no correlated motion between loops 1 and 7. However, in the hA3C R44Q/R45H/A144S variant, there is a region of increased correlation between these two regions, indicating that the A144S mutation affects the motion of nearby loop 7, which in turn is more correlated with the motion of loop 1 (Figure E.11). In the rhA3C, the opposite effect is expected, where the open conformation time of rhA3C Q44R/H45R/S144A loop 1 is increased compared to rhA3C WT. This is consistent with previous observations that more time in an open loop 1 state correlates with increases in deamination activity and is consistent with our observation of increased specific activity for rhA3C Q44R/H45R/S144A (Figure 6.7, Figure 6.9C, and Figure E.9).[257]

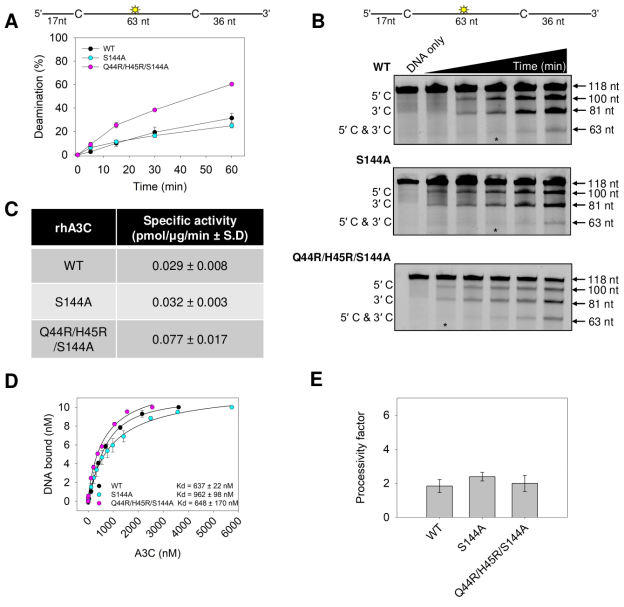


FIGURE 6.9. **An S144A amino acid change in rhA3C increases deamination activity, but not processivity.**

6.3.7 Efficient HIV Restriction by rhA3C when Dimerization is Mediated by Amino Acid 144

To test if the rhA3C S144A-mediated dimer form enabled HIV restriction, we conducted a single-cycle infectivity assay. We found that the rhA3C Q44R/H45R/S144A (Figure 6.10A, 33% infectivity) could restrict HIV at an equivalent level to hA3C S188I (Figure 6.10A, 24% infectivity). This was not due to increased encapsidation as it was encapsidated into HIV virions at an equivalent or slightly lesser amount that hA3C S188I (Figure 6.10B). Although the rhA3C S144A could dimerize, it could not restrict HIV, consistent with no increase in catalytic activity (Figure 6.8, Figure 6.9C, and Figure 6.10A). This result was interesting since both the S144A and M115N induced dimerization in combination with the Q44R/H45R mutations, but the rhA3C Q44R/H45R/S144A had a higher level of HIV restriction activity (Figure 6.10A, 33% infectivity and Figure 6.4A, 53% infectivity). We hypothesized that this simply could be because the Q44R/H45R/S144A had a more dominant dimer peak than Q44R/H45R/M115N (compare Figure 6.2A-B and Figure 6.8A-B). Alternatively, since the *in vitro* specific activities were similar, the difference could be due to a larger change of loop 1 for rhA3C Q44R/H45R/S144A compared to rhA3C Q44R/H45R/M115N (Figure 6.7 and Figure E.9). Since the rhA3C Q44R/H45R/S144A processivity was less than rhA3C Q44R/H45R/M115N (compare Figure 6.3D and Figure 6.9D) this led us to hypothesize that the rhA3C Q44R/H45R/S144A restricts HIV by a predominantly deamination-independent mode (mediated by nucleic acid binding) and rhA3C Q44R/H45R/M115N a deamination-dependent mode (mediated by deamination).

We tested the deamination-dependent mode by determining the level of G→A mutations in the coding strand of the integrated proviral DNA. We tested hA3C S188I, rhA3C WT and the two rhA3C mutants, Q44R/H45R/S144A and Q44R/H45R/M115N. The hA3C S188I had the highest level of G→A mutations (Table 6.1, 6.54 G→A mutations/kb). The majority mutations were in the GA→AA context (5'TC on the (-)DNA), which is the commonly preferred context for hA3C (Table 6.1). Consistent with lower activity against HIV (Figure 6.10A), the rhA3C WT had the lowest level of G→A mutations (Table 6.1, 2.15

mutations/kb). For rhA3C WT there was an approximately equal amount of mutations with a GG→AG context (5'CC on the (-)DNA) and GA→AA context, indicating that the rhA3C active site has a more relaxed sequence preference. The GG→AG context is primarily associated with A3G.[276] The rhA3C Q44R/H45R/S144A induced 1.5-fold less mutations than the rhA3C Q44R/H45R/M115N (Table 6.1, 3.20 and 4.67 G→A mutations/kb, respectively) consistent with rhA3C Q44R/H45R/S144A being 2-fold less processive than rhA3C Q44R/H45R/M115N.

To further explore how the rhA3C Q44R/H45R/S144A could restrict HIV more than rhA3C Q44R/H45R/M115N we also determined the level of proviral DNA integration. A3s can inhibit reverse transcriptase activity, which results in less completed proviral DNA synthesis, and less integration.[248,249] The rhA3C WT and rhA3C Q44R/H45R/M115N did not decrease proviral DNA integration (Figure 6.10C). However, hA3C S188I and rhA3C Q44R/H45R/S144A did decrease proviral DNA integration (Figure 6.10C). The rhA3C Q44R/H45R/S144A allowed only 56% of the total proviral DNA to integrate, relative to the no A3 condition (Figure 6.10C). These data show that rhA3C Q44R/H45R/S144A restricts HIV using both deamination -dependent and -independent modes of restriction which are more effective than only the deamination-dependent mode used by rhA3C Q44R/H45R/M115N. Since the ssDNA binding affinities for rhA3C Q44R/H45R/S144A and Q44R/H45R/M115N were similar (Figure 6.3E and Figure 6.9D), these data suggest that the rhA3C Q44R/H45R/-S144A changes to loop 1 dynamics, increased dimerization, or both features enabled more deamination-independent restriction than rhA3C Q44R/H45R/M115N.

6.3.8 Evolutionary Dynamics of Key Residues Involved in A3C Dimerization

Since amino acids at residues 44, 45, 115, and 144 were all found to be important in the gain of dimerization and restriction activity of rhA3C, we examined the evolution of these residues over primate evolution. None of these residues corresponds to those that were previously reported to be under positive selection.[223] Nonetheless, they do vary in Old World Primates and in Hominoids. While both the rhesus and the crab-eating macaque encode the QHMS at residues 44, 45, 115, and 144, respectively (Figure 6.11), the Northern

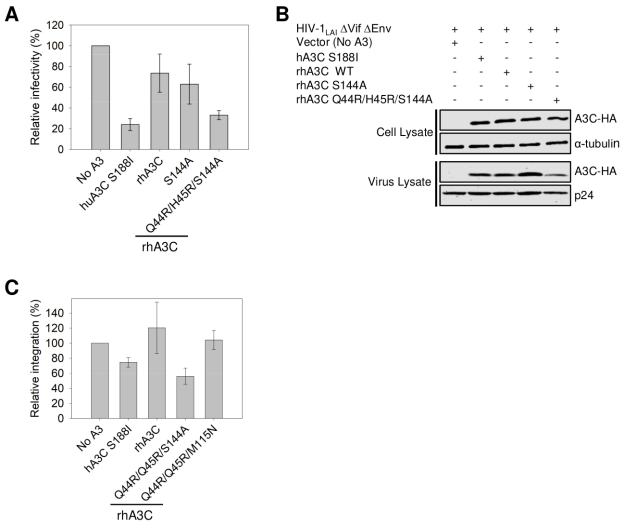


FIGURE 6.10. **Amino acids 144 in combination with 44 and 45 in rhA3C enable HIV restriction ability. (A)** Infectivity was measured by β-galactosidase expression driven by the HIV-1 5 LTR from TZM-bl cells infected with VSV-G pseudotyped HIV ΔVif ΔEnv that was produced in the absence or presence of 3xHA tagged hA3C S188I, rhA3C WT, and rhA3C mutants S144A and Q44R/H45R/S144A. Results normalized to the no A3 condition are shown with error bars representing the Standard Deviation of the mean calculated from three independent experiments. **(B)** Immunoblotting for the HA tag was used to detect A3C enzymes expressed in cells and encapsidated into HIV ΔVif ΔEnv pseudotyped virions. The cell lysate and virion loading controls were α-tubulin and p24, respectively. **(C)** The relative amount of proviral DNA integration in infected HEK293T cells in the presence of hA3C S188I, rhA3C WT, and rhA3C mutants Q44R/H45R/S144A and Q44R/H45R/M115N in comparison to the No A3 condition was determined by qPCR. Error bars represent the standard deviation of the mean calculated from at two independent experiments.

pig tailed macaque encodes a cysteine (C) at residue 45, while the nearest outgroup, the Baboon, encodes an arginine (R) at residue 45, which is the residue found in hominids at that

position (Figure 6.11). We used the sequences of 17 Old World Monkeys and Hominoids as well as one New World Monkey (for purposes of rooting the tree) to reconstruct the ancestral amino acids at each of these positions (Figure 6.11). We find evidence that the glutamine (Q) at amino acid 44 which is unfavorable for dimerization of rhA3C is, in fact, a derived trait as the ancestral amino acid at position 44 at both the root of the rhesus/baboon/drill common ancestor as well as the common ancestor that includes the African green monkeys (Sabaeus, Grivet, Tantalus, Vervet) is a histidine (H) at position 44. These results suggest that loss of dimerization of rhA3C is an ancient event but may not have occurred by the same mechanism in all lineages.

6.4. Discussion

Throughout the evolution of a host restriction factor, the protein must be able to retain activity against different viral pathogens, which may require compensatory mutations over time to either keep up with evolution of the initial virus or to counter-act new viral transmissions within the species.[236] In this study, we investigated the impact of evolutionary changes on a specific A3 enzyme, rhA3C. Our work highlights the importance of an integrated strategy to identify relevant functional interactions in enzymes. In particular, it describes a cycle of experiment and theory that allowed us to efficiently identify key positions from a combinatorial web of potential interactions. This methodology is transferrable and warrants its application to study other members of the APOBEC family as our analysis showed that the evolutionary signals for dimerization seem to be preserved across multiple organisms and family members with distinct functions. This methodology would also be useful in other protein systems.

Determination of activity against lentiviruses for A3 enzymes is multifactorial. First, it requires viral encapsidation, which occurs for all A3C orthologues tested thus far (Figure 6.4).[223,235] After that, there is a requirement to deaminate cytidines in lentiviral (-)DNA or inhibit reverse transcriptase.[277] Despite rhA3C inducing 2 mutations/kb, approximately 20 mutations in the total genome, the infectivity of the HIV was not significantly decreased (Figure 6.10 and Table 6.1). Since the mutations are stochastic, there may be some genomes

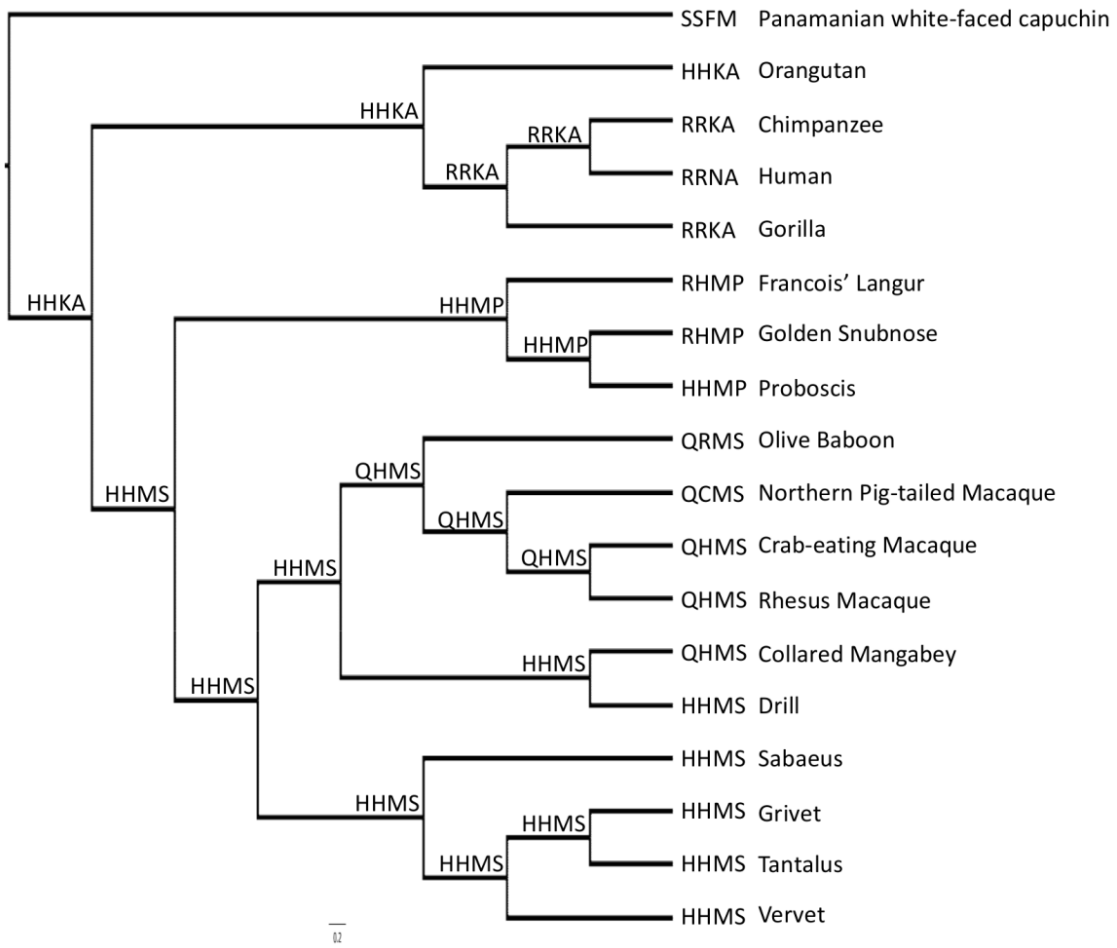


FIGURE 6.11. **Ancestral reconstruction of residues involved in rhA3C dimerization.** A phylogeny depicting residues crucial for A3C dimerization, including reconstructed ancestral residues. The species included are on the right of the phylogenetic tree of A3C sequences as well as the four amino acids at residues 44, 45, 115, and 144, respectively. The identities of the amino acids at each of these positions as calculated by FASTML are shown at each node of the tree.

with many mutations, some with none, and some with mutations that are not inactivating. Therefore, the more mutations, the better for ensuring inactivation. Increased processivity can achieve this, as evidenced by the 2-fold increase in mutations by the more processive rhA3C Q44R/H45R/M115N mutant, but this was still not enough for robust HIV restriction

A3 enzyme	Base Pairs Sequenced	Total G→A mutations	Total GG→AG mutations	Total GA→AA mutations	G → A Mutations per kb	GG→AG mutations per kb	GA→AA mutations per kb
No A3	8715	4	2	2	0.46	0.23	0.23
hA3C S188I	8715	57	15	41	6.54	1.72	4.70
rhA3C WT	6972	15	6	7	2.15	0.86	1.00
rhA3C Q44R/ H45R/ S144A	8134	26	12	11	3.20	1.48	1.35
rhA3C Q44R/ H45R/ M115N	8134	38	20	15	4.67	2.46	1.84

TABLE 6.1. Analysis of A3-induced mutagenesis of *pol* gene from integrated HIVΔVif.

(Figure 6.4 and Table 6.1). What was needed was a deamination independent restriction of reverse transcriptase in combination with inducing mutagenesis to enable more robust restriction as for rhA3C Q44R/H45R/S144A (Figure 6.10 and Table 6.1).

However, the rhA3C Q44R/H45R/S144A results seemly go against numerous studies showing a direct correlation between processivity and mutation frequency in multiple A3 enzymes, including hominid A3Cs.[215,235,251,278] Using an *in vitro* system to study the effect of processivity on mutations during reverse transcription, it was shown that A3A, a non-processive enzyme, could introduce a similar number of mutations to A3G, a highly processive enzyme, but not at the same locations along the ssDNA.[279] The processivity was needed for mutations to accumulate in ssDNA regions rapidly lost to replication and dsDNA formation. The deaminations of A3A were achieved by a quasi-processive search involving multiple on and off interactions with the ssDNA. In this process time is lost, but deaminations still occur. This may be occurring with rhA3C Q44R/H45R/S144A.

Also interesting was that for rhA3C, unlike hominid A3Cs, the dimer did not always

result in a processive enzyme. The rhA3C Q44R/H45R/S144A had equal specific activity to rhA3C Q44R/H45R/M115N but by a seemingly different mechanism than processivity (Figure 6.3 and Figure 6.9). In addition, the rhA3C S144A single mutant could dimerize, but had no increase in activity (Figure 6.8 and Figure 6.9). The rhA3C Q44R/H45R/S144A is predicted to have an altered loop 1 that is able to increase the deamination activity (Figure 6.7 and Figure E.9). Loop 1 has been identified to be a gate-like structure that controls access of the ssDNA to the active site.[274,275] A3A that has an open loop 1 conformation and displays high specific activity with low processivity.[280] In contrast, A3B and a related family member Activation Induced Cytidine deaminase (AID) have closed loop structures.[274,275] However, A3B and AID are processive and still achieve similar activity to A3A but are more selective for which cytidines are deaminated or under which types of conditions, e.g., A3B is more active when ssDNA is in excess to the enzyme.[281] Our data suggest that the rhA3C Q44R/H45R/S144A has increased activity in an A3A-like manner (Figure 6.9 and Figure 6.10). Further, the rhA3C S144A mutant demonstrates that dimerization was necessary, but not sufficient, since arginines at residues 44 and 45 were also required for the increase in specific activity and restriction activity (Figure 6.9 and Figure 6.10). The reason for why the processive rhA3C Q44R/H45R/M115N does not have an increase in anti-viral activity similar to hA3C S188I may simply be due to the lack of a stable dimer, which would be needed in the absence of the loop 1 alteration (Figure 6.2). Loop 1 has previously been shown to regulate activity of hA3C.[257] The hA3C was shown to have an amino acid pair 25W/26E that decreased activity compared to a hA3C 25R/26K mutant.[257] Our MD simulations data suggest that alterations in the protein structure from a 144A substitution changed the dynamics of loop 1 indirectly, which also resulted in increased specific activity, in a 44R/45R background (Figure 6.7, Figure 6.9, and Figure E.9).

Interestingly, the majority of positive selection in A3C has taken place outside of the interaction motif with Vif, which suggests an evolution of enzyme activity, rather than avoidance of the lentiviral antagonist, providing a unique model to study restriction factor evolution.[223] Here we examined the residues needed for activity against HIV in rhA3C in

their evolutionary context in other OWMs. We found that none of the analyzed OWM A3Cs contained the right amino acid combinations for robust activity, although there were considerable changes. The rhA3C sequences at the four key amino acids is ancient, but is a derived trait from the common ancestor of Old World Monkeys (Figure 6.11). Other OWMs also had other changes to one or two sites. This perhaps indicates that different selective pressures were on OWM A3C from a non-lentivirus pathogen. This may be a retroelement, such as LINE-1, since hA3C can restrict LINE-1 similarly to hA3C S188I, indicating that there is no requirement for dimerization.[223] Alternatively, A3C may act in concert with other A3s. For example, in humans, A3G and A3F have been found to hetero-oligomerize and this increases the activity of both enzymes.[282] This hetero-oligomerization has been understudied and perhaps A3C acts with another A3 and has greater activity. This would perhaps explain the different requirements for dimerization in comparison to hA3C although rhA3C and hA3C use the same general interface. Finally, it is possible that A3C, like A3H, has lost activity during evolution in primates,[283] perhaps because of selection against the deleterious effects of the enzyme, or because its activity has been usurped by other A3 enzymes.

6.5. Conclusions

The data show that rhA3C activity can be enhanced through dimerization that either increases processivity or more robustly through dimerization that causes an alteration of loop 1 conformation and dynamics. This is important for understanding the biochemical basis of activity of A3C and other A3 family members and for using it as a tool to predict anti-lentiviral activity in other OWMs. The OWM A3C has evidence of positive selection both within and outside the Vif binding region, suggesting that it has antiviral activity.[223] However, the fixation of the I188 in rhA3C and likely other OWMs did not impart anti-lentiviral activity, perhaps due to other compensatory mutations being made at the sites identified here, which were needed to combat other pathogens. Altogether, the data provide an in depth structure-function analysis of rhA3C and suggest that the rhA3C viral targets of restriction have yet to be thoroughly identified.

CHAPTER 7

INVESTIGATION OF DRIVERS FOR PREFERENTIAL CARBOXY-*S*-ADENOSYL-L-METHIONINE SYNTHESIS BY M.MPEI VARIANTS

Manuscript in preparation

7.1. Introduction

Nucleotide analogues (NAs) or non-canonical nucleotides are useful in an increasing variety of ways, acting as antiviral agents, medical imaging agents, mutagens, and even as a means of producing synthetic life.[284,285] A number of NAs are used today as antiviral agents against diseases such as HIV (didanosine[286] , abacavir[287] , zidovudine[288] , zalcitabine[289]), Hepatitis B (emtricitabine[290] , entecavir[291] , telbivudine[292]), and the SARS-CoV2 novel coronavirus that causes COVID-19 (remdesivir[293]). There are also NAs that act as antibiotics, such as muraymycin,[294] and a plethora of NAs used as anti-cancer drugs.[295] NAs can act as mutagens, with a common example found in 5-bromouracil (5-BrU), which spontaneously isomerises to pair with guanine instead of adenine.[296] Interestingly, 5-BrU can cause GC–to–AT mutations and AT–to–GC transversions, depending on experimental protocols.[297] The use of NAs as medical imaging agents is well established, as many NAs act as fluorophores.[298–300] NAs as fluorophores are extremely useful as they are more able to pass through cellular membranes for intracellular imaging, and can be easily metabolized and removed from the body afterward.[295,300] Some NAs have been introduced as "unnatural base pairs", providing additional options for genetic coding of synthetic biology.[284,285] There are some naturally occurring modified or non-canonical nucleobases.[301] These modifications relate to epigenetic effects, including gene silencing and regulation of expression, transcription, and replication.[302–304] One common modification is the methylation of cytidine to 5-methylcytidine, the latter of which accounts for 4% of cytidines[303,305]

The use of enzymes in the production of pharmaceutically relevant compounds is of particular interest, as they can be produced quickly and at commercially viable scales.[306–308] Furthermore, rational design principles may be applied to develop new nucleotide analogues

that may be produced via *in vitro* enzymatic catalysis.[309,310] Rational design principles also extends to enzyme structure and function.[311–313] Enzymes are modified to provide increased activity,[314] altered selectivity or preference for a given substrate,[315] or even an entire reaction mechanism.[316]

Methyltransferases (MTases) are a category of enzymes that catalyse the methylation of nucleotides, peptides, and small molecule natural products.[317,318] Widespread loss of methylation is recognized as a marker of oncogenesis and the progression of certain cancers.[319] Direct peptide methylation can affect intermolecular interactions between proteins and substrates.[320] MTases may function as part of gene regulation mechanisms and to prevent degradation by enzymes.[321] Some MTases have been used for directed labeling of biomolecules with *S*-adenosyl-L-methionine (SAM) and SAM-analogues.[322] They are also used in pharmaceutical synthesis, with enzyme catalysis playing an increasing role in bringing novel drugs to market. Additionally, MTases have previously been modified to use different cofactors through directed evolution combined with enzyme promiscuity, allowing further modification of DNA-based drugs.[323,324]

M.MpeI is a SAM-dependent, CpG-specific DNA MTase found in *Mycoplasma penetrans* (Figure 7.1).[325] The normal function of M.Mpe1 involves methylation of an incoming cytidine with SAM, producing a 5mC nucleotide and *S*-adenosylhomocysteine (SAH). Recent work has reported that M.MpeI can synthesis the modified nucleotide 5-carboxymethylcytosine (5cxmC) which is formed as a trace byproduct of cytidine methylation by M.MpeI with the secondary metabolite carboxy-*S*-adenosyl-L-methionine (CxSAM) in place of the normal reaction cofactor, SAM.[326] In that work, Wang *et al.* reported an N374K mutation that improves selectivity for the CxSAM cosubstrate.[326]

A deeper understanding of how the N374K variant selects for the CxSAM co-substrate can aid to devise additional mutation profiles for experimental study. Here, we explore the mechanism driving the altered selectivity of the N374K mutation, and use rational design to explore possible mutations that may further drive the preference of M.MpeI for CxSAM over SAM using a combined theoretical and experimental approach.

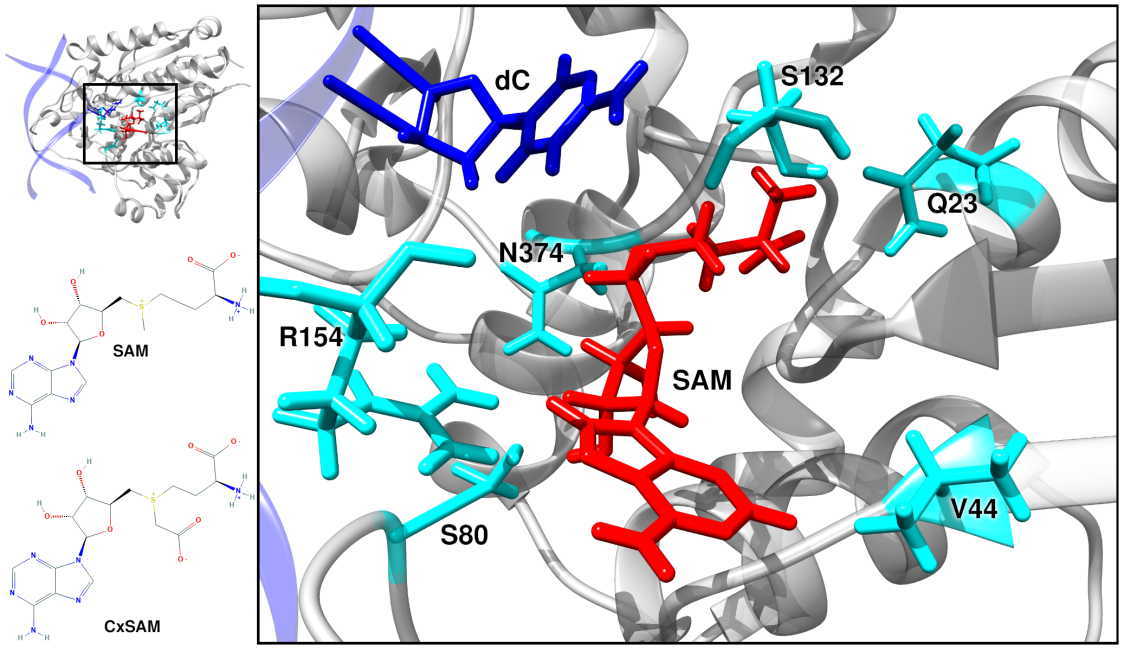


FIGURE 7.1. Active site of CpG-specific DNA methyltransferase M.Mpe1 with standard and novel ligands.

7.2. Computational Methods

Five protein variants were prepared from the original crystal structure (PDB: 4DKJ[325]): wildtype (WT), N374K, E45G, E45D, and E45D/N374K. For each variant, systems were prepared with the SAM or CxSAM co-substrates, resulting in a total of ten unique systems. The protonation states of amino acid side chains were determined using the H++ server.[216,260,261] Custom forcefield parameters were generated for the SAM and CxSAM ligands using the PyRED server.[327–330] These forcefields were used along with FF14SB for the protein, TIP3P for the water, counterions, and metal ions, and OL15 for the DNA. Each system was neutralized to zero net charge using K+ counterions, and the system was solvated in TIP3P water with a minimum distance of 12 Å between the protein surface and the edge of the periodic boundary resulting in a periodic unit cell measuring 71Å x 64Å x 51Å.

Each system was minimized over 50 steps of steepest descent followed by 450 steps of conjugate gradient at 10K with the protein, ligand, and DNA frozen to allow the density

of the solvent box to equilibrate. The system was heated from 10K to 300K over 20 stages, with each stage taking 12.5 ps. After heating, the restraints on all non-solvent molecules were gradually removed from 100 kcal mol^{-1} $Å^{-2}$ in 10 stages. Equilibration and production were run at a temperature of 300 K and 1.0 atm using a Berendsen thermostat in an NVT ensemble. The nonbonded cutoff was set to 8.0 Å and a smooth particle-mesh Ewald method was used for long-range Coulomb interactions.[159] Each system was equilibrated for 50 ns before production to ensure system stability. Production was run with a 1 fs timestep for 250 ns using the pmemd.cuda module in AMBER18.[151] All bonds involving hydrogen atoms were constrained using the SHAKE algorithm. All systems were simulated in triplicate for a total of 750 ns of MD sampling per ligand/variant combination.

Correlated motion, hydrogen bonding interactions, root-mean-squared deviation (RMSD) and fluctuation (RMSF), normal modes, and distances were calculated using cpptraj.[331] Energy decomposition analysis (EDA) was performed using AMBER-EDA.[264]

7.3. Computational Results

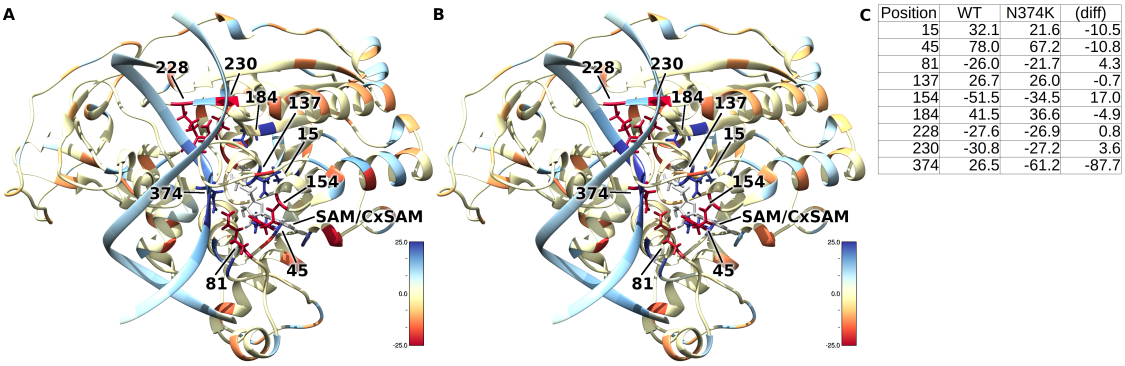
7.3.1 Mutations and Ligands Do Not Impact Dynamic Motion of the System

Each of the ten systems (WT, E45G, E45D, N374K, or E45D/N374K with SAM or CxSAM) was evaluated to determine if the structure or dynamic motion of the protein was affected by either point mutation or substrate. Multiple metrics of dynamic motion (RMSF, normal modes, motion cross-correlation, principle component analysis) were compared across all variant-ligand combinations to ensure protein stability. In all cases the RMSD was constant over the span of the production time, indicating the systems were stable with respect to the crystal structure and did not undergo any large conformational changes (see Figure F.1). RMSF of all system was consistent across multiple replicates of all variants and ligands (see Figure F.2). The normal modes, while different in magnitudes between replicate trajectories, maintained similar overall motion profiles, indicating that the essential dynamics remain relatively unchanged across all systems (see Figures F.3 and F.4). Correlated motion of residue pairs also show similar patterns across systems (see Figure F.5). Taken together,

these data indicate that the changing ligands and point mutations did not significantly impact the dynamic motion of the protein nor its overall structure.

N374K mutation exhibits preference for CxSAM

Energy decomposition analysis revealed that wildtype (WT) M.MpeI favors SAM over CxSAM by -133.9 kcal mol^{-1}. The N374K variant also favors SAM over CxSAM, however this difference is reduced to only -59.5 kcal mol^{-1}. The N374K/SAM interaction differs from the WT/SAM interaction by +28.9 kcal mol^{-1} (less favorable), while the N374K/CxSAM interaction differs from the WT/CxSAM interaction by -45.6 kcal mol^{-1} (more favorable). The specific interactions driving these differences are highlighted in Figure 7.2. At position 374 in the WT, the interaction favors SAM by 26.5 kcal mol^{-1}. In the N374K variant, this selectivity changes to CxSAM by -61.2 kcal mol^{-1}, a total change in interaction energy of 87.7 kcal mol^{-1}.



Position	WT	N374K	(diff)
15	32.1	21.6	-10.5
45	78.0	67.2	-10.8
81	-26.0	-21.7	4.3
137	26.7	26.0	-0.7
154	-51.5	-34.5	17.0
184	41.5	36.6	-4.9
228	-27.6	-26.9	0.8
230	-30.8	-27.2	3.6
374	26.5	-61.2	-87.7

FIGURE 7.2. **a)** Difference in interaction energies between WT with CxSAM against SAM baseline. Residues highlighted in blue (red) interact more favorably with SAM (CxSAM). **b)** Difference in interaction energies between N374K with CxSAM with respect to SAM. **c)** Interaction energy differences between highlighted residues and co-substrates in kcal mol^{-1}. Positive (negative) values indicate the residue at that position interacts more favorably with SAM (CxSAM).

7.3.2 E45 is a Good Candidate for Mutagenic Study

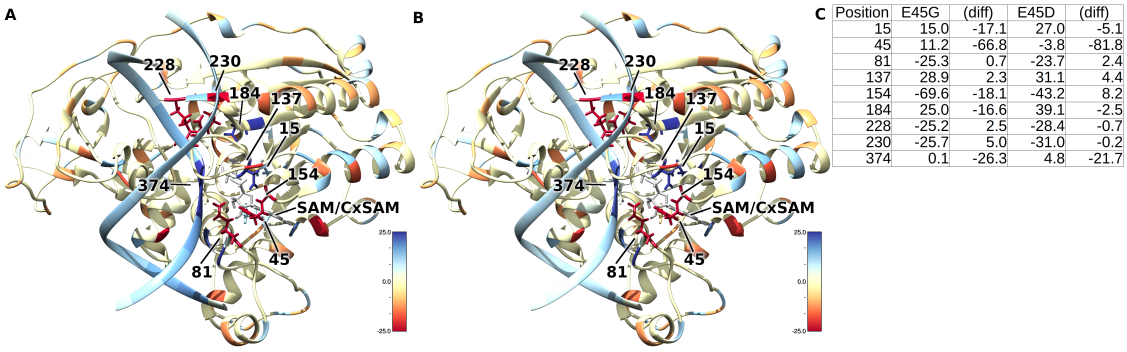
The largest differences in interaction energies between the WT and SAM/CxSAM co-

substrates is reported in Figure 7.2**a**,**c**. The largest of these differences in the WT occurred at position E45, corresponding to a change of 78.0 kcal mol^{-1} less favorable in its interactions with the CxSAM compared to SAM. To investigate this position, we ran simulations with E45G to remove the negative charge and the bulk of the side chain, and E45D to shorten the side chain without removing the negative charge.

The E45G system showed a reduction in the ΔE between position 45 and the co-substrates, however the total difference in interaction energy between the protein/DNA complex and the co-substrates is also less favorable for both when compared to other systems (see Figure 7.4). These results suggest that while E45G does alter selectivity between substrates, it would also be less likely to interact with either substrate at all when compared to other variants investigates. Conversely, the E45D variant shows an 81.8 kcal mol^{-1} more favorable selectivity for CxSAM over SAM, with comparatively minor changes at other residues examined from the WT. Interestingly, the E45D variant also resulted in a significant improvement at position N374 (see Figure 7.3). The overall interaction energy between the protein/DNA complex and the SAM co-substrate was similar to that of the N374K variant, while the interaction of the E45D variant with CxSAM was 17.4 kcal mol^{-1} more favorable than the N374K variant. Taken together, the E45D and E45G data indicate that the charged residue at this position is important for binding, however the smaller side chain of the aspartate variant appears to provide more accessible volume to accommodate the larger CxSAM substrate.

7.3.3 E45D and N374K Provide Additive Effects on Substrate Selectivity

The interaction energies between the identified residues and the co-substrates are shifted further towards CxSAM selectivity in the E45D/N374K double mutant (see Figure 7.4). Interestingly, the interaction between D45 and the co-substrate is favored towards CxSAM in the double mutant than in the single mutant, and the same is true for the K374 position. This suggests that the improvement of selectivity of these two positions is likely due to many body effects, as the double mutant includes both a positive and negative charge on opposite sides of the CxSAM cosubstrate, leading to greater stabilization. The total



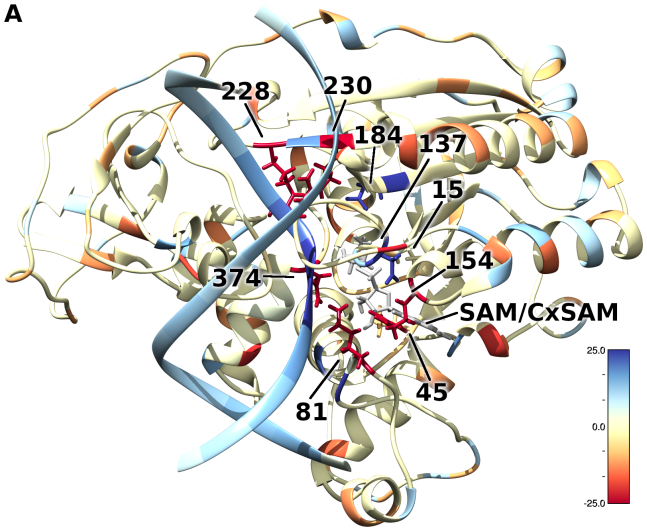
C

Position	E45G	(diff)	E45D	(diff)
15	15.0	-17.1	27.0	-5.1
45	11.2	-66.8	-3.8	-81.8
81	-25.3	0.7	-23.7	2.4
137	28.9	2.3	31.1	4.4
154	-69.6	-18.1	-43.2	8.2
184	25.0	-16.6	39.1	-2.5
228	-25.2	2.5	-28.4	-0.7
230	-25.7	5.0	-31.0	-0.2
374	0.1	-26.3	4.8	-21.7

FIGURE 7.3. **a)** Difference in interaction energies between E45G with CxSAM against SAM baseline. Residues highlighted in blue (red) interact more favorably with SAM (CxSAM). **b)** Difference in interaction energies between E45D with CxSAM with respect to SAM. **c)** Interaction energy differences between highlighted residues and cosubstrates in kcal mol^{-1}. Positive (negative) values indicate the residue at that position interacts more favorably with SAM (CxSAM).

interaction energies between the protein/DNA complex and the co-substrates also selects for CxSAM over SAM by 44.2 kcal mol^{-1}, and this interaction energy is at the same magnitude as that of the WT and single mutants with the SAM cosubstrate. These data indicate that the E45D/N374K double mutant should preferentially bind CxSAM over SAM to approximately the same degree as the WT and single variants with SAM.

A



B

Position	E45D/N374K	(diff)
15	-0.9	-33.0
45	-7.3	-85.4
81	-29.1	-3.1
137	33.2	6.6
154	-43.0	8.4
184	42.3	0.8
228	-32.6	-4.9
230	-34.4	-3.7
374	-76.3	-102.8

C

	SAM	CxSAM	Selectivity
WT	-222.2	-88.3	-133.9
N374K	-193.3	-133.9	-59.5
E45D	-196.3	-151.2	-45.1
E45G	-103.5	-82.0	-21.6
E45D/N374K	-155.5	-199.6	44.2

FIGURE 7.4. **a)** Difference in interaction energies between E45D/N374K with CxSAM against SAM baseline. Residues highlighted in blue (red) interact more favorably with SAM (CxSAM). **b)** Interaction energy differences between highlighted residues and co-substrates in kcal mol^{-1}. Positive (negative) values indicate the residue at that position interacts more favorably with SAM (CxSAM). **c)** Total interaction energies in kcal mol^{-1} between protein/DNA complex and SAM/CxSAM co-substrates.

7.4. Conclusions

These data suggest that the N374K mutation results in a shift in interaction energies that increasingly favors the CxSAM cosubstrate. The interaction energies observed between the two cosubstrates and the variants under investigation show that the E45D mutation similarly favors CxSAM, and that the two mutations are additive, resulting in an E45D/N374K double mutation that preferentially acts on CxSAM over SAM, pushing the promiscuous enzyme M.MpeI from primarily using SAM to primarily selecting for CxSAM. These mutations are also useful because they do not significantly impact the structure or dynamic motion of the protein or the interaction energies between the DNA and protein.

CHAPTER 8

IMPLEMENTATION OF MINIMUM FREE ENERGY PATH OPTIMIZATION ALGORITHM IN LICHEM

8.1. Introduction

Quantum mechanical/molecular mechanical (QM/MM) simulations of reaction mechanisms are an important tool in studies of enzyme catalysis, especially in investigations of how reaction rates and substrate binding are affected by mutations or other environmental effects. LICHEM is a software package that calls external software packages for both QM (Gaussian, etc) and MM (TINKER, etc) calculations and transitions between them to obtain optimized reaction paths and energies for large enzymatic systems.[106] The current single point and reaction path optimization algorithms minimize the geometry on the potential energy surface. Optimizations in this way can produce geometries that are "ideal", but not necessarily accurate with respect to reaction mechanisms. The inclusion of free energy calculations in the algorithm can produce results that account for the average forces on each atom over a period of sampling time.

The Hamiltonian of the QM atoms in a molecular dynamics trajectory can be calculated using Equation 8.1. This accounts for the different conformations of the atoms in the MM region during dynamics and computes an average field. The complete free energy of the system is calculated using equation 8.2. For the complete derivation, see reference **296**.

$$H = H_{QM}(r_{QM}) + \frac{1}{N}\sum_{n=1}^{N}\sum_{j\in MM}\sum_{i\in QM}\frac{q_j}{|r_i - r_{n,j}|} \tag{8.1}$$

$$A_0 = -\frac{1}{\beta}\ln\left[\int e^{-\beta A_0(r_{QM})} dr_{QM}^{M}\right] \tag{8.2}$$

Minimum free energy perturbation (MFEP) calculations allow us to include thermodynamic properties of a system at non-zero temperatures and account for the multiple states

that a system may occupy over the course of a reaction.[332] These additional effects can improve the calculated reaction energies and barriers for enzymatic reactions in solution.[60,333] The calculation of MFEP involves molecular dynamics simulations to obtain sufficient sampling to calculate an average field of forces and energies in which the reaction mechanism may be calculated. The computational workflow is shown in Figure 8.1.

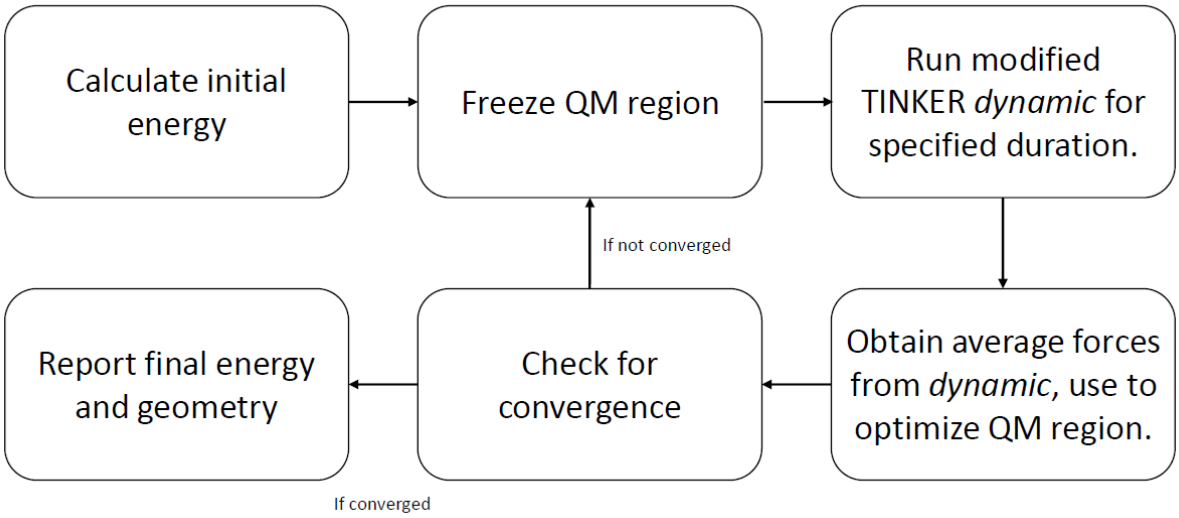


FIGURE 8.1. Algorithm for MFEP implementation in LICHEM

8.2. Software Updates

The TINKER source code for *dynamic* and *gradient* was modified to calculate, store, and output the average forces on each atom for use in LICHEM. LICHEM was modified to accept new MFEP-related keywords to allow users to change the number of dynamic steps, timestep, ensemble (NPT, NVE, NVT, NPH), temperature, and pressure. The *use_MFEP* keyword also allows the use of the modified *dynamic* in place of standard *minimize* calls during normal OPT, DFP, and QSM calculation types. These programs were then tested to ensure the modifications do not otherwise impact the functionality.

8.3. Testing Methods

The diaspartate system found in the LICHEM tutorial package was used to separately test individual components of the MFEP workflow and to obtain calculation timings. This

system contains two aspartic acid molecules solvated in 31,981 water molecules in a cube measuring 98.6 Å on a side, totalling 95,993 atoms. The QM region selected consists of the side chains of the aspartic acids up to the α-carbon, with the peptide backbone atoms included in the MM region (see Figure 8.2).

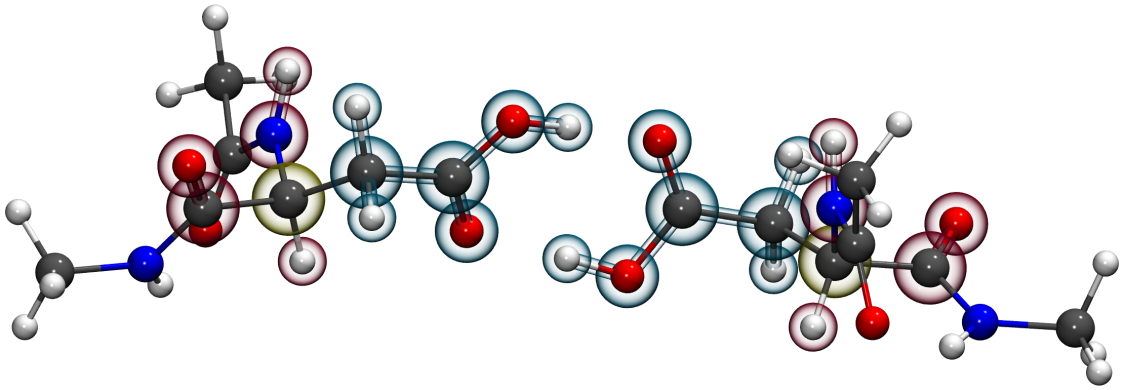


FIGURE 8.2. Diaspartate test system. Atoms highlighted in blue are in the QM region, yellow are pseudobond atoms, and red are boundary atoms.

An unmodified TINKER 7.1 package was first used to ensure the test system behaved as expected in both *minimize* and *dynamic*. The minimization was performed to an RMS gradient threshold criteria of 0.1Å. The unmodified *dynamic* and modified *dynamic* were both run for 1000 dynamic steps using a 1.0fs timestep in an NPT ensemble at 300K and 1.0 bar. These criteria were selected to ensure that the modifications to *dynamic* did not seriously impact calculation time or accuracy. All test calculations were run on UNT's Cruntch3 using processing nodes that each provided 20 compute cores.

I ran small tests with the LICHEM MFEP implementation with dynamics set to run for 2.0fs timesteps for 25, 50, 100, and 200 steps using a long-range electrostatic correction cutoff of 7 Å for nonbonded interactions, in an NPT ensemble at 300K and 1.0 bar, and using particle mesh ewald.[159] The QM atoms were represented with the PBE0 functional with the 6-31+G(d,p) basis set.[334] The diaspartate test system includes a product and reactant structure, both of which were optimized using the MFEP method at all four listed dynamic durations. The calculated energies of each tested system at all four timestep options and the

	Timesteps	Free Energy (kcal/mol)
Reactant	25	-1775253.1
	50	-1775253.1
	100	-1775253.4
	200	-1775253.4
	0 (Potential)	-1157865.1_4
Product	25	-1775253.7
	50	-1775253.7
	100	-1775253.7
	200	-1775253.7
	0 (Potential)	-1157865.1_8
Reaction Free Energy (ΔG)	25	-0.6
	50	-0.6
	100	-0.3
	200	-0.3
Reaction Potential Energy (ΔE)	0	-0.04

TABLE 8.1. Free energies for reactant and product of diaspartate double proton transfer reaction test system and reaction energies based on timing settings tested.

resulting reaction free energies are listed in Table 8.1. For comparison, potential energies from calculations performed in previous work are also included.

These data indicate that the energies remain consistent with different durations of dynamic timings. They also show that the reaction energies are within an acceptable error, with the 25– and 50–timestep tests showing a reaction free energy of -0.6 kcal mol^{-1}, decreasing to -0.3 kcal mol^{-1} for the 100– and 200–timestep tests. The previously reported reaction potential energy is -0.04 kcal mol^{-1}. This indicates that when thermal and entropic contributions from MFEP are included, the product is slightly more stabilized.

While these tests have not been run to a length of time to obtain acceptable amounts of sampling data for useful comparisons to experimental data, they establish that the implementation of minimum free energy in LICHEM does work for single optimizations and without significant increase in computational costs.

CHAPTER 9

CONCLUSIONS AND FUTURE WORK

The APOBEC3 and M.MpeI projects together demonstrate that protein mutations may be predicted to serve a purpose or goal, whether to rescue lost functionality,[2] impact dimerization,[5] or to modify the selectivity and preference of a promiscuous enzyme. Predictive modeling is highly useful as it can help guide experimental work with improved specificity. APOBEC3H may be modified to restore lost anti-cancer activity due to a mutation. APOBEC3C dimerization is affected by the interplay of just a few amino acids that differ between the human and rhesus variants. M.MpeI methyltransferase can be modified using rational design principles to cause it to favor a nonstandard substrate over its usual substrate, and the effects of individual mutations are additive. The method development I did with the combined conserved evolution and electron localization resulted in an improved approach to the selection of QM region atoms in large enzymatic systems, which can enable calculations with higher levels of theory or lower computational costs, improving accuracy and throughput of results in these studies. The inorganic complex project demonstrates that low barrier hydrogen bonds may be formed and tuned on a tridentate pincer ligand with modified substituent groups.

Going forward, the APOBEC3H projects would ideally be followed up with a QM/MM study of the reaction mechanism and how it is affected by the mutations at positions 117 and 121. The reaction mechanism is not yet fully understood. The cytidine deaminase of *E. coli* consists of a different active site configuration, but is still Zinc-dependent and uses a glutamate-coordinated water to catalyze the reation.[335,336] The deaminase in yeast is a much closer homolog to APOBEC3s, with the same active site configuration. Both of these reaction mechanisms have been investigated with QM/MM methods, and serve as an excellent point of comparison for a future APOBEC3H QM/MM study. The reaction free energies for the wildtype, K121E, K117E, and K117E/K121E variants would provide additional information as to the complex way in which APOBEC3H mutations can affect cancer

pathogenesis.

The M.MpeI methyltransferase enzyme could be similarly investigated using QM/MM methods, and may also benefit from simulations with other SAM-analogues. The promiscuous enzyme could serve to be a useful tool in the drug design toolkit if it can be further tuned to preferentially act upon a variety of analogues to functionalize nucleobases. An initial study of known SAM-analogues in the active site, followed by proposed new analogues based on desired modified nucleobases, could provide many new avenues of research.

The MFEP implementation in LICHEM needs further testing with a wider array of systems, as well as testing on a complete reaction path using the QSM method. Additionally, a comparison between the results of MFEP using AMBER point charges and the results using AMOEBA would demonstrate that the implementation produces consistent results, and that AMOEBA provides further improvement to the accuracy from MFEP calculations. I will perform tests on the diaspartate test system up to 1ns of dynamic simulation time for the reactant and product and each of the beads in the QSM reaction path. Additionally, I will use the 4-oxalocrotonate tautomerase (4OT) system described in Chapter 3 to serve as a test of MFEP on an enzymatic system rather than a simple small molecule reaction.

APPENDIX A

CHARACTERIZING HYDROGEN-BOND INTERACTIONS IN PYRAZINETETRACARBOXAMIDE COMPLEXES: INSIGHTS FROM EXPERIMENTAL AND QUANTUM TOPOLOGICAL ANALYSES

Experimental Methods

For full experimental methods, see text of Ref **4**.

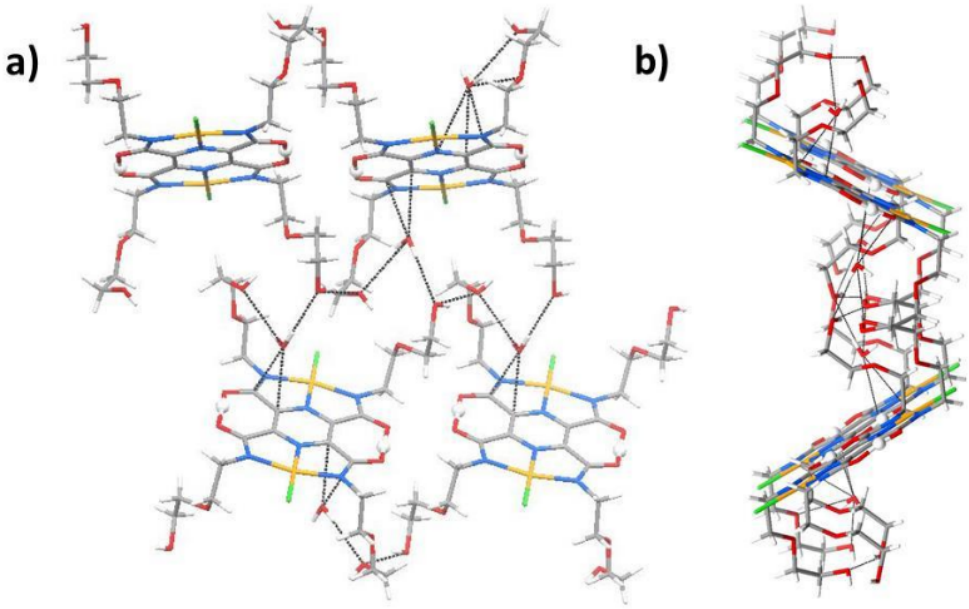
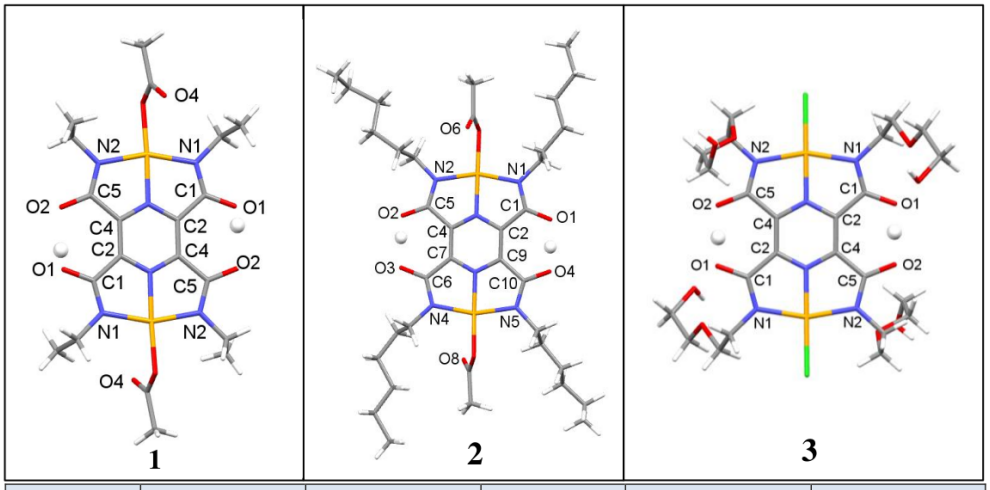


FIGURE A.1. Views of tetramers for the palladium chloride complex **3**



atoms	distance	atoms	distance	atoms	distance
O1-C1	1.281(5)	O1-C1	1.280(5)	O1-C1	1.282(4)
O2-C5	1.275(5)	O2-C5	1.280(5)	O2-C5	1.277(4)
O2'-C5'	1.275(5)	O3-C6	1.278(4)	O2'-C5'	1.277(4)
O1'-C1'	1.281(5)	O4-C10	1.288(5)	O1'-C1'	1.282(4)
N1-C1	1.297(5)	N1-C1	1.299(5)	N1-C1	1.305(4)
N2-C5	1.301(5)	N2-C5	1.297(5)	N2-C5	1.301(4)
N2'-C5'	1.301(5)	N4-C6	1.309(5)	N2'-C5'	1.301(4)
N1'-C1'	1.297(5)	N5-C10	1.292(5)	N1'-C1'	1.305(4)

TABLE A.1. Selected interatomic distances [Å] for complexes, with perspective views directly above appropriate columns to show atom numbering schemes for **1**, **2**, and **3**. Primes (') indicate atoms related by inversion center.

Computational Methods

The wavefunctions for ELF, NCI, and QTAIM calculations were obtained with the Gaussian 09 softtware package, performing single point calculations using the ωB97XD functional with a mixed basis set (LanL2DZ for Pd atoms, and 6-311+G(d,p) for C, H, O, and N atoms). ELF calculations were computed with the ToPMoD software package using a cubic grid of 200 au with a step size of 0.1 au. For the NCI calculations, the NCIPLOT software package was used, with a cutoff of 0.5. Both ELF and NCI surfaces were rendered with

VMD. For the QTAIM calculation, the AIMALL software package was used.

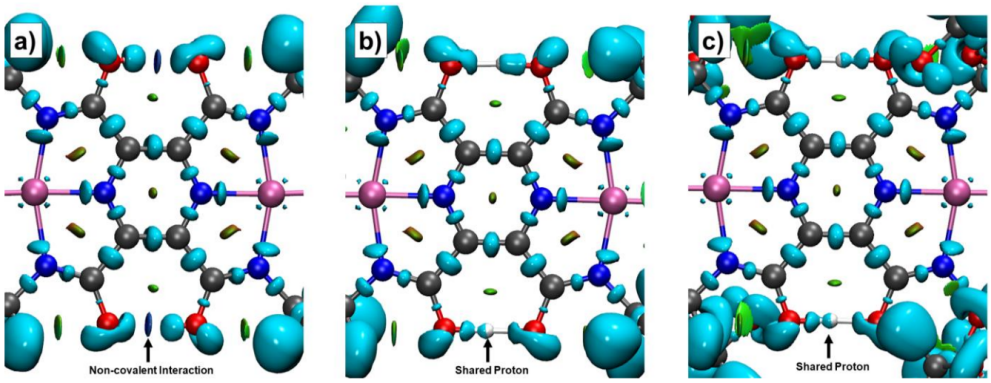


FIGURE A.2. Close-up of the combined ELF/NCI analysis for the two H-bond regions for complexes **1** (a), **2** (b) and **3** (c). ELF topologies are shown as solid cyan surfaces, and NCI regions are rendered as solid surfaces. In NCI analysis, as per convention, red surfaces indicate strong repulsion, blue surfaces denote strong attraction and green surfaces indicate weak interactions. Isosurface cutoff for ELF= 0.85. Isosurface cutoff for NCI= 0.5, and data is plotted in the range -0.05 $<$ sign(λ2)ρ $<$ 0.05 a.u.

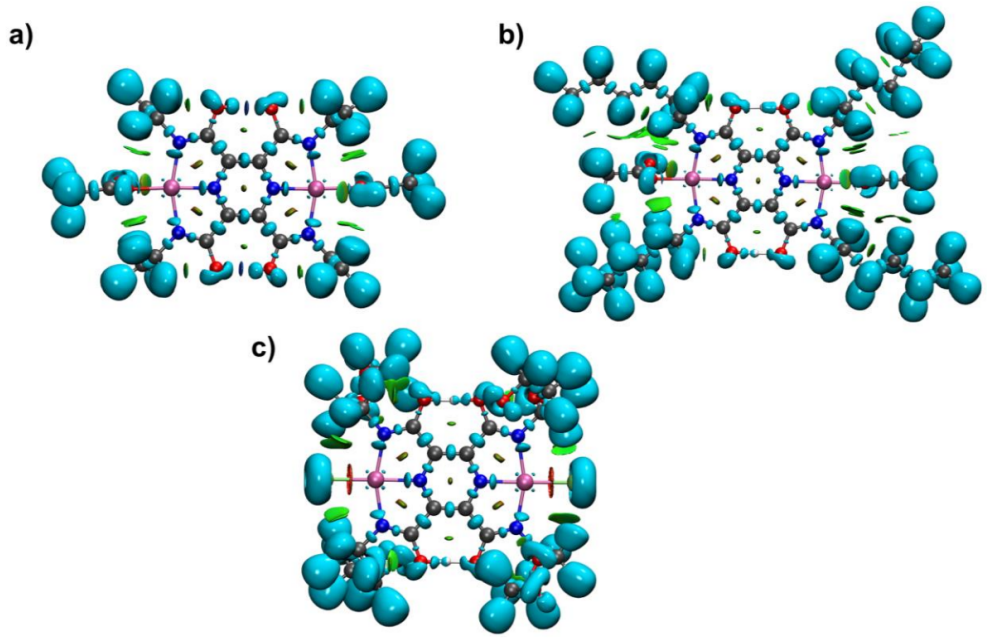


FIGURE A.3. Combined ELF/NCI topological representation of complexes **1** (a), **2** (b) and **3** (c). ELF basins appear as solid cyan surfaces, and NCI regions are represented as solid surfaces, where red surface indicates strong repulsion, blue surface strong attraction and green surface relatively weak interactions. Isosurface cutoff for ELF=0.85. Isosurface cutoff for NCI=0.5, and data is plotted in the range -0.05 < sign(λ2)ρ < 0.05 a.u.

ELF population and distributed multipole analyses for selected basins (Tables A.2-A.4). Population and multipole analysis for complex **3** shows that the first and second polar moments for the lone pair basins that are directly interacting with the H atom are significantly larger than the ones pointing away from the H bond interaction region. This is an indication of the strong polarization on the O atoms induced by the LBHB. In addition, the population analysis shows that each of the oxygen atoms that share the H atom for complex **3** have a similar electron population (O2=4.03,1.62; O1=3.73,1.98), on the other hand complex **1** does not (O2=2.81,2.99; O1=3.86). This further suggests the presence of a covalent bond between H1-O1 for complex **1**. For complex **2**, the populations for the monosynaptic and disynaptic basins are similar to the ones for complex **3** and complex **1**,

respectively. The 1st and 2nd polar moments for each of the largest valence basins V(O2) and V(O1) decreased between 1-0.6 a.u, respectively, while for the valence basin V(O2') for both polar moments decreased ∼0.6 a.u., and for the V(O1'), increased ∼0.3 a.u. In addition, the population and distributed multipole analyses for the V(O2,C5) and V(O1,C1) basins for complexes **1** and **3**, indicates that when an LBHB is present, the population, as well the first and second polar moments, for these disynaptic basins are roughly the same. On the other hand, if a covalent bond is present, their population is slightly smaller.

Basin	**Population**	**1st Polar moment (a.u.)**	**2nd Polar moment (a.u.)**
V(H1,O1)	2.15	1.055	0.43
V(O1)	3.86	2.723	1.978
V(O1,C1)	1.77	0.121	0.294
V(O2)	2.81	1.092	1.032
V(O2)	2.99	1.262	0.937
V(O2,C5)	1.88	0.137	0.673
V(H2,O1')	2.14	1.048	0.416
V(O1')	3.88	2.761	2.009
V(O1',C1')	1.76	0.116	0.298
V(O2')	2.96	1.245	0.948
V(O2')	2.82	1.099	1.027
V(O2',C5')	1.88	0.143	0.677

TABLE A.2. Electron population, 1st and 2nd polar moments for the specific basins for complex **1**. The two numbers per column correspond to each long pair associated with the valence electrons for each O atom.

Basin	Population	1st Polar moment (a.u.)	2nd Polar moment (a.u.)
V(H1)	0.54	0.072	0.125
V(O2)	3.60	1.832	1.040
V(O2)	1.91	0.685	0.272
V(O2,C5)	2.16	0.178	0.241
V(O3)	3.44	1.674	0.868
V(O3)	2.10	0.562	0.376
V(O3,C6)	2.18	0.305	0.347
V(H2,O4)	2.34	1.921	1.696
V(O4)	3.74	2.100	1.425
V(O4,C10)	2.24	0.388	0.349
V(O1)	3.29	1.667	0.730
V(O1)	2.44	0.692	0.544
V(O1,C1)	2.05	0.234	0.393

TABLE A.3. Electron population, 1st and 2nd polar moments for the specific basins for complex **2**. The two numbers per column correspond to each long pair associated with the valence electrons for each O atom.

Basin	Population	1st Polar moment (a.u.)	2nd Polar moment (a.u.)
V(H1)	0.35	0.039	0.092
V(O2)	4.03	2.855	1.99
V(O2)	1.62	0.395	0.216
V(O2,C5)	1.84	0.091	0.443
V(O1)	3.73	2.314	1.561
V(O1)	1.98	0.553	0.335
V(O1,C1)	1.85	0.081	0.49
V(H2)	0.35	0.039	0.092
V(O2')	4.03	2.855	1.99
V(O2')	1.62	0.395	0.216
V(O2',C5')	1.84	0.091	0.443
V(O1')	3.73	2.312	1.561
V(O1')	1.97	0.548	0.337
V(O1',C1')	1.85	0.08	0.495

TABLE A.4. Electron population, 1st and 2nd polar moments for the specific basins for complex **3**. The two numbers per column correspond to each long pair associated with the valence electrons for each O atom.

QT-AIM analysis of complexes 1-3 (Tables A.5-A.7). Covalent bonds (shared interactions) possess negative values for $\nabla^2\rho(\mathbf{r})$ and small values for $\rho(\mathbf{r})$. Closed shell

interactions, like HBs, also have small $\rho(\mathbf{r})$ but positive $\nabla^2\rho(\mathbf{r})$ values. For complex **1** (Table A.5), the $\nabla^2\rho(\mathbf{r})$ values for the BCP H1-O1 and H2-O1' were negative and positive for O1-H2 and O2'-H2. For complex **3** (Table A.7), all the $\nabla^2\rho(\mathbf{r})$ values were negative. For complex **2** (Table A.6), the value of $\nabla^2\rho(\mathbf{r})$ for BCP corresponding to H1-O2, O3-H1 and H2-O4 is negative, while the one corresponding to O1-H2 is positive. These results are consistent with the ELF analysis and indicate that complex **1** exhibit a short HB while **3** exhibits LBHBs; for complex 2, one of the HB corresponds to an LBHB while the other to a short HB.

Bond Critical point (BCP)	$\rho(\mathbf{r})\left(e\,a_0^{-3}\right)$	$\nabla^2\rho(\mathbf{r})\left(e\,a_0^{-5}\right)$
O1-H1	6.28E-01	-7.63
H1-O2	4.89E-02	2.04E-01
H2-O1'	6.27E-01	-7.59
O2'-H2	4.88E-02	2.04E-01

TABLE A.5. Bond critical point electron densities, $\rho(\mathbf{r})$, and Laplacians, $\nabla^2\rho(\mathbf{r})$, for HBs in complex **1**.

Bond Critical point (BCP)	$\rho(\mathbf{r})\left(e a_0^{-3}\right)$	$\nabla^2\rho(\mathbf{r})\left(e a_0^{-5}\right)$
H1-O2	1.35E-01	-1.74E-02
O3-H1	2.07E-01	-7.44E-01
H2-O4	2.74E-01	-1.67
O1-H2	1.02E-01	1.40E-01

TABLE A.6. Bond critical point electron densities, $\rho(\mathbf{r})$, and Laplacians, $\nabla^2\rho(\mathbf{r})$, for HBs in complex **2**.

Bond Critical point (BCP)	$\rho(\mathbf{r})\left(e a_0^{-3}\right)$	$\nabla^2\rho(\mathbf{r})\left(e a_0^{-5}\right)$
H1-O2	2.13E-01	-7.98E-01
O1-H1	1.36E-01	-8.27E-03
O2'-H2	2.13E-01	-7.98E-01
H2-O1'	1.36E-01	-8.26E-03

TABLE A.7. Bond critical point electron densities, $\rho(\mathbf{r})$, and Laplacians, $\nabla^2\rho(\mathbf{r})$, for HBs in complex **3**.

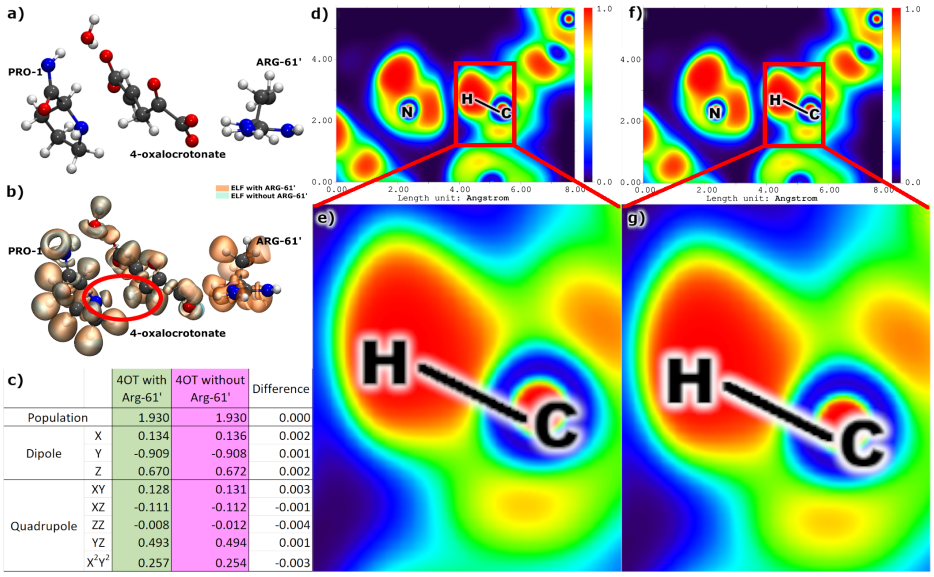
Mulliken for the HBs of interest (Table A.8). For complex **1**, the charge difference between the O atoms participating in the HB, O1 and O2, is large (approx. 0.17), with a smaller charge observed on the oxygen bonded that is bound to the H. For complex **3**, the charge difference between O1 and O2 is not significant. For complex **2**, the charge difference for O2 and O3 has a similar trend as complex **3**, while for O1 and O4 the charge difference is comparable with complex **1**. In all complexes, the charge difference between H1 and H2 was negligible.

	Atom	*O1*	*H1*	*O2*	*O1'*	*H2*	*O2'*
Complex **1**	*Mulliken Charge*	-0.2508	0.3924	-0.4214	-0.2508	0.3926	-0.4210
	Atom	*O3*	*H1*	*O2*	*O1*	*H2*	*O4*
Complex **2**	*Mulliken Charge*	-0.4297	0.5713	-0.4349	-0.3586	0.5410	-0.4271
	Atom	*O1*	*H1*	*O2*	*O1'*	*H2*	*O2'*
Complex **3**	*Mulliken Charge*	-0.3852	0.5077	-0.3021	-0.3852	0.5077	-0.3020

TABLE A.8. Mulliken charges for HBs in complexes **1**, **2**, and **3**

APPENDIX B

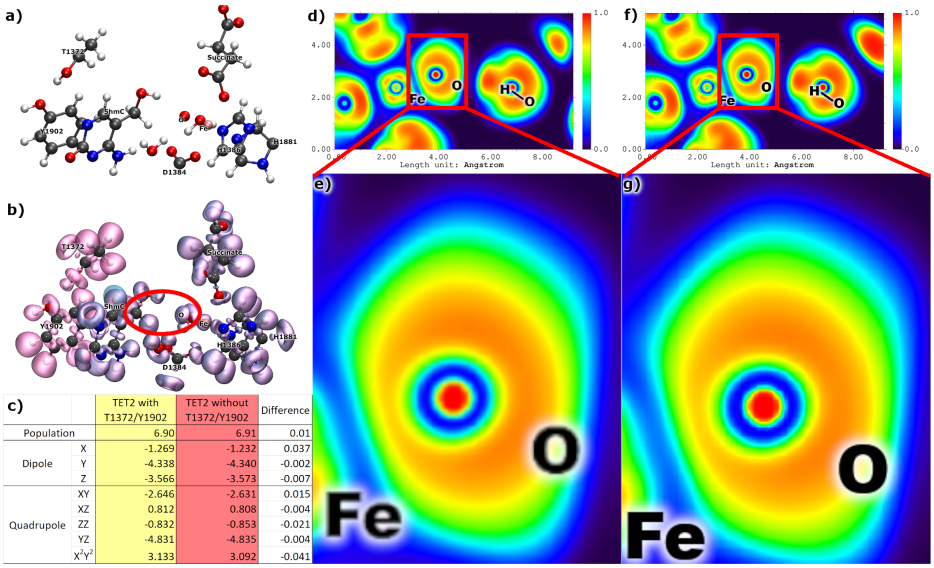
COMBINING EVOLUTIONARY CONSERVATION AND QUANTUM TOPOLOGICAL ANALYSES TO DETERMINE QM SUBSYSTEMS FOR BIOMOLECULAR QM/MM SIMULATIONS



c)

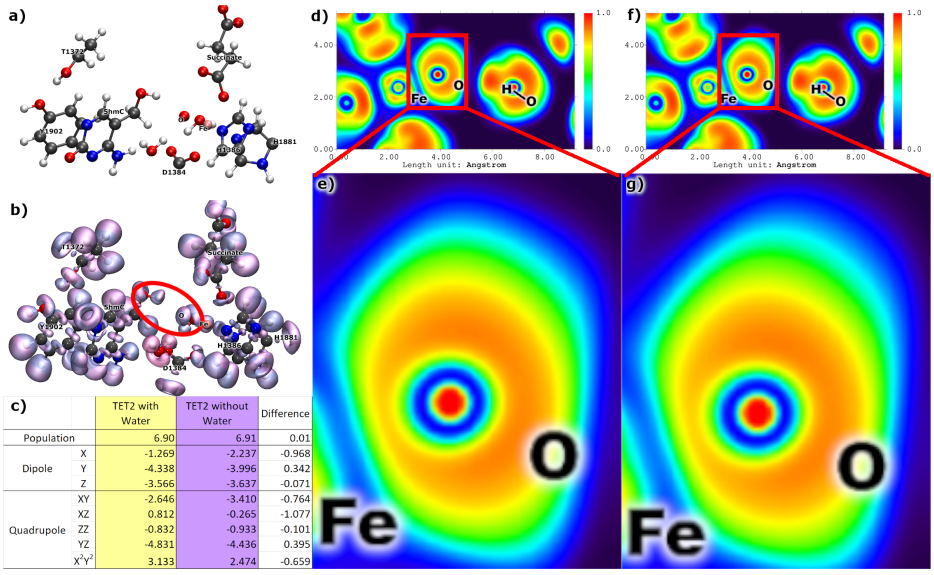
		4OT with Arg-61'	4OT without Arg-61'	Difference
Population		1.930	1.930	0.000
Dipole	X	0.134	0.136	0.002
	Y	-0.909	-0.908	0.001
	Z	0.670	0.672	0.002
Quadrupole	XY	0.128	0.131	0.003
	XZ	-0.111	-0.112	-0.001
	ZZ	-0.008	-0.012	-0.004
	YZ	0.493	0.494	0.001
	X^2Y^2	0.257	0.254	-0.003

FIGURE B.1. **a)** 4OT active site with considered residues. **b)** ELF for 4OT with and without Arg-61' included in the QM region. **c)** Comparison of the multipolar decomposition for the basin of the transfered proton. **d)** ELF heatmap of 4OT with Arg-61' on the plane passing through the reactive bond (C-H) and the reaction-involved proline nitrogen. **e)** Closeup of the ELF heatmap for the reactive bond region without Arg-61' in QM subsystem. **f)** ELF heatmap of 4OT without Arg-61' on the plane passing through the reactive bond (C-H) and the reaction-involved proline nitrogen. **g)** Closeup of the ELF heatmap for the reactive bond with Arg-61' in QM subsystem.



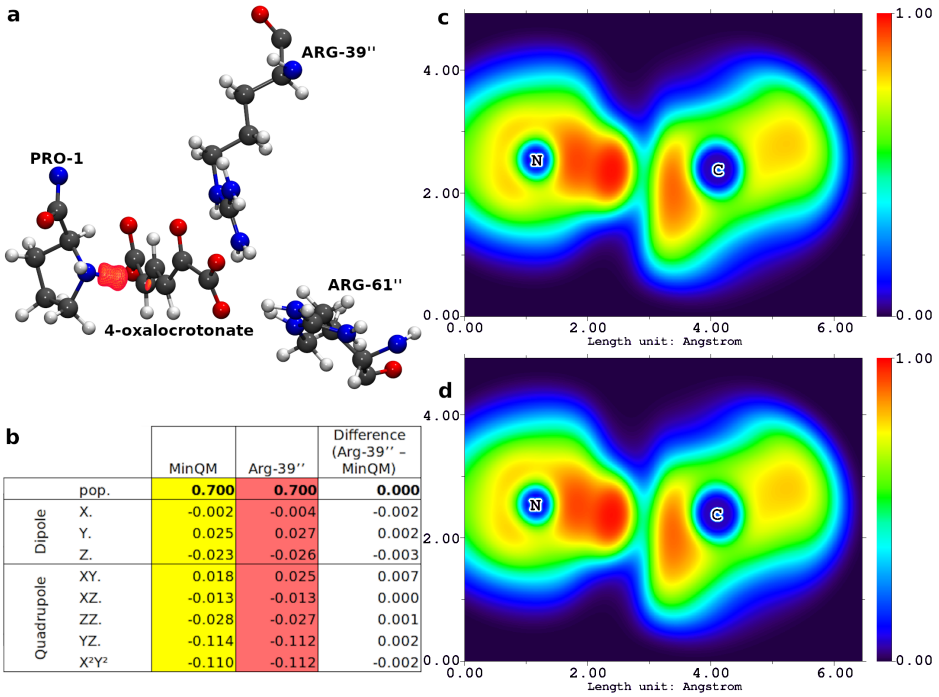
c)		TET2 with T1372/Y1902	TET2 without T1372/Y1902	Difference
Population		6.90	6.91	0.01
Dipole	X	-1.269	-1.232	0.037
	Y	-4.338	-4.340	-0.002
	Z	-3.566	-3.573	-0.007
Quadrupole	XY	-2.646	-2.631	0.015
	XZ	0.812	0.808	-0.004
	ZZ	-0.832	-0.853	-0.021
	YZ	-4.831	-4.835	-0.004
	X^2Y^2	3.133	3.092	-0.041

FIGURE B.2. **a)** QM region for TET2 system. **b)** ELF for TET2 with and without T1372 and Y1902 in the QM region. **c)** Comparison of electron localization with and without these two residues. **d)** ELF heatmap of TET2 without additional amino acids on the plane passing through the reactive iron-oxyl bond and the 5' hydroxyl group of the cytidine substrate. **e)** Closeup of iron-oxyl without additional amino acids in QM region. **f)** ELF heatmap of TET2 with T1372 and Y1902 on the plane passing through the reactive iron-oxyl bond and the 5' hydroxyl group of the cytidine substrate. **g)** Closeup of iron-oxyl with additional amino acids in QM region.



c)		TET2 with Water	TET2 without Water	Difference
Population		6.90	6.91	0.01
Dipole	X	-1.269	-2.237	-0.968
	Y	-4.338	-3.996	0.342
	Z	-3.566	-3.637	-0.071
Quadrupole	XY	-2.646	-3.410	-0.764
	XZ	0.812	-0.265	-1.077
	ZZ	-0.832	-0.933	-0.101
	YZ	-4.831	-4.436	0.395
	X^2Y^2	3.133	2.474	-0.659

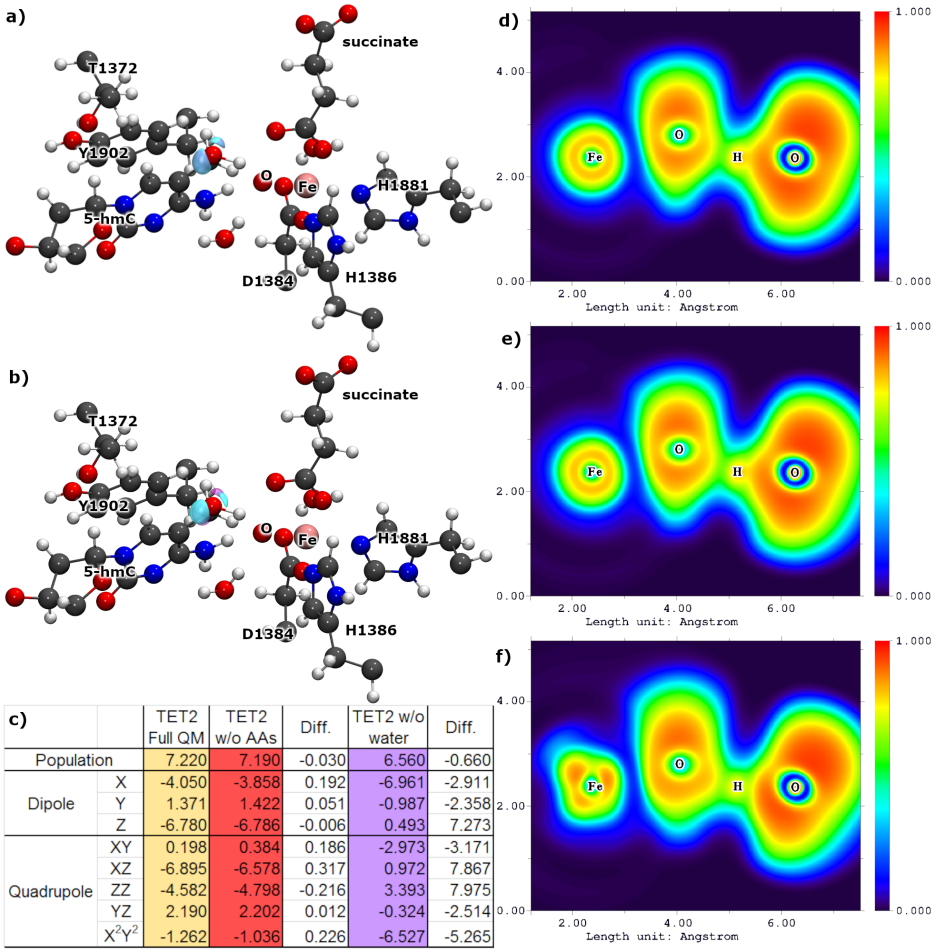
FIGURE B.3. **a)** QM region for TET2 system. **b)** ELF for TET2 with and without second shell water included in the QM region. **c)** Comparison of electron localization with and without water. **d)** ELF heatmap of TET2 without second shell water on the plane passing through the reactive iron-oxyl bond and the 5' hydroxyl group of the cytidine substrate. **e)** Closeup of iron-oxyl without water in the QM region. **f)** ELF heatmap of TET2 with second shell water on the plane passing through the reactive iron-oxyl bond and the 5' hydroxyl group of the cytidine substrate. **g)** Closeup of iron-oxyl with water in the QM region.



b

		MinQM	Arg-39''	Difference (Arg-39'' – MinQM)
	pop.	**0.700**	**0.700**	**0.000**
Dipole	X.	-0.002	-0.004	-0.002
	Y.	0.025	0.027	0.002
	Z.	-0.023	-0.026	-0.003
Quadrupole	XY.	0.018	0.025	0.007
	XZ.	-0.013	-0.013	0.000
	ZZ.	-0.028	-0.027	0.001
	YZ.	-0.114	-0.112	0.002
	X^2Y^2	-0.110	-0.112	-0.002

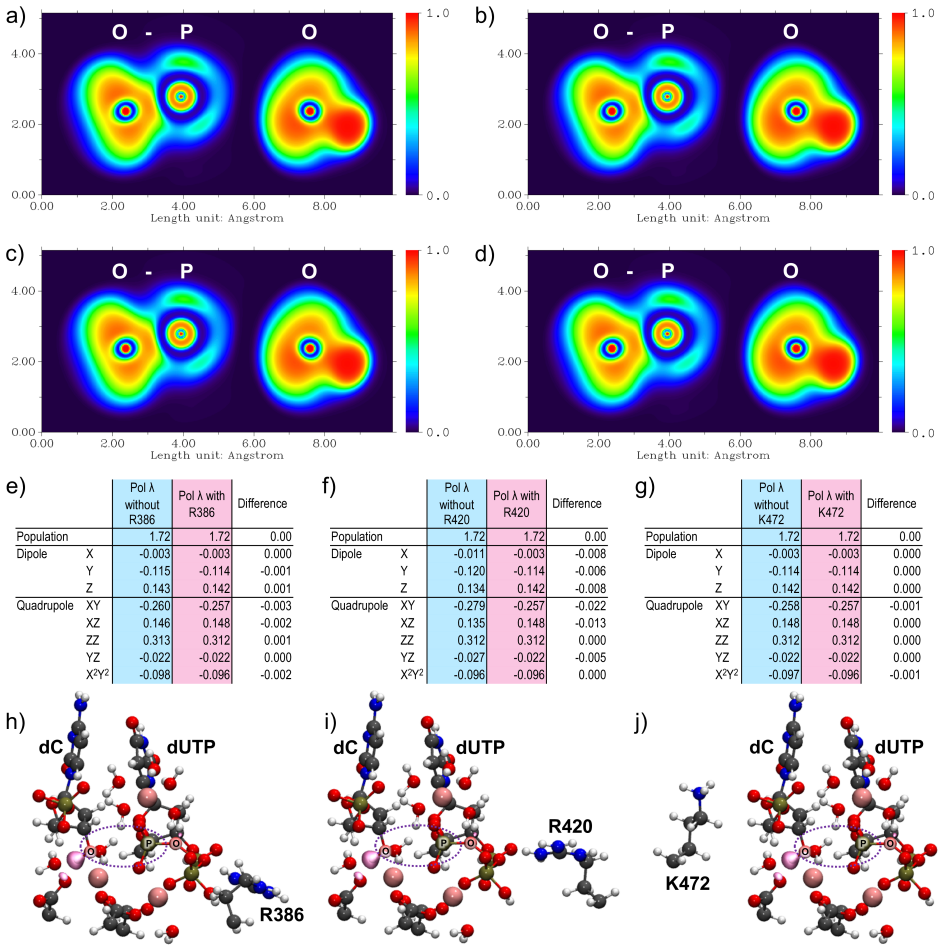
FIGURE B.4. **a)** ELF of truncated wavefunction around reactive atoms in 4OT transition state with (wireframe) and without (solid transparency) Arg-39" in the QM region, **b)** comparison of electron localization and multipolar moments with and without Arg-39", **c)** ELF heatmap through plane of reactive atoms with Arg-39" represented as point charges in the MM region. **d)** ELF heatmap through plane of reactive atoms C,H, and N with Arg-39" represented in the QM region,



c)

		TET2 Full QM	TET2 w/o AAs	Diff.	TET2 w/o water	Diff.
Population		7.220	7.190	-0.030	6.560	-0.660
Dipole	X	-4.050	-3.858	0.192	-6.961	-2.911
	Y	1.371	1.422	0.051	-0.987	-2.358
	Z	-6.780	-6.786	-0.006	0.493	7.273
Quadrupole	XY	0.198	0.384	0.186	-2.973	-3.171
	XZ	-6.895	-6.578	0.317	0.972	7.867
	ZZ	-4.582	-4.798	-0.216	3.393	7.975
	YZ	2.190	2.202	0.012	-0.324	-2.514
	X^2Y^2	-1.262	-1.036	0.226	-6.527	-5.265

FIGURE B.5. **a)** ELF of truncated wavefunction around reactive atoms in TET2 transition state with and without T1372 and Y1902 in the QM region, **b)** ELF of truncated wavefunction around reactive atoms in TET2 transition state with and without second shell water in the QM region. **c)** comparison of electron localization and multipolar moments of the iron-oxyl electronic basin for full QM region, QM without T1372 and Y1902, and QM without second shell water, **d)** ELF heatmap through plane of reactive atoms Fe, O, and H with the full QM region, **e)** ELF heatmap through plane of reactive atoms with T1372 and Y1902 represented as point charges in the MM region, **f)** ELF heatmap through plane of reactive atoms with second shell water represented as point charges in the MM region.



e)

		Pol λ without R386	Pol λ with R386	Difference
Population		1.72	1.72	0.00
Dipole	X	-0.003	-0.003	0.000
	Y	-0.115	-0.114	-0.001
	Z	0.143	0.142	0.001
Quadrupole	XY	-0.260	-0.257	-0.003
	XZ	0.146	0.148	-0.002
	ZZ	0.313	0.312	0.001
	YZ	-0.022	-0.022	0.000
	X^2Y^2	-0.098	-0.096	-0.002

f)

		Pol λ without R420	Pol λ with R420	Difference
Population		1.72	1.72	0.00
Dipole	X	-0.011	-0.003	-0.008
	Y	-0.120	-0.114	-0.006
	Z	0.134	0.142	-0.008
Quadrupole	XY	-0.279	-0.257	-0.022
	XZ	0.135	0.148	-0.013
	ZZ	0.312	0.312	0.000
	YZ	-0.027	-0.022	-0.005
	X^2Y^2	-0.096	-0.096	0.000

g)

		Pol λ without K472	Pol λ with K472	Difference
Population		1.72	1.72	0.00
Dipole	X	-0.003	-0.003	0.000
	Y	-0.114	-0.114	0.000
	Z	0.142	0.142	0.000
Quadrupole	XY	-0.258	-0.257	-0.001
	XZ	0.148	0.148	0.000
	ZZ	0.312	0.312	0.000
	YZ	-0.022	-0.022	0.000
	X^2Y^2	-0.097	-0.096	-0.001

FIGURE B.6. 2D ELF visualization of Polλ **a)** for the original QM region, **b)** for the QM region with R386, **c)** for the QM region with R420, and **d)** for the QM region with K472. Comparison of electron localization without (blue) and with (pink) **e)** R386, **f)** R420, and **g)** K472 on the plane passing through the reactive phosphate bond. ELF for Polλ (blue) with **h)** R386, **i)** R420, and **j)** K472 in the QM region.

APPENDIX C

COMPUTATIONAL INVESTIGATION OF APOBEC3H SUBSTRATE ORIENTATION AND SELECTIVITY

C.1. Results for Substrate Binding Orientation Studies

Raw data plots for orientation and recognition simulations.

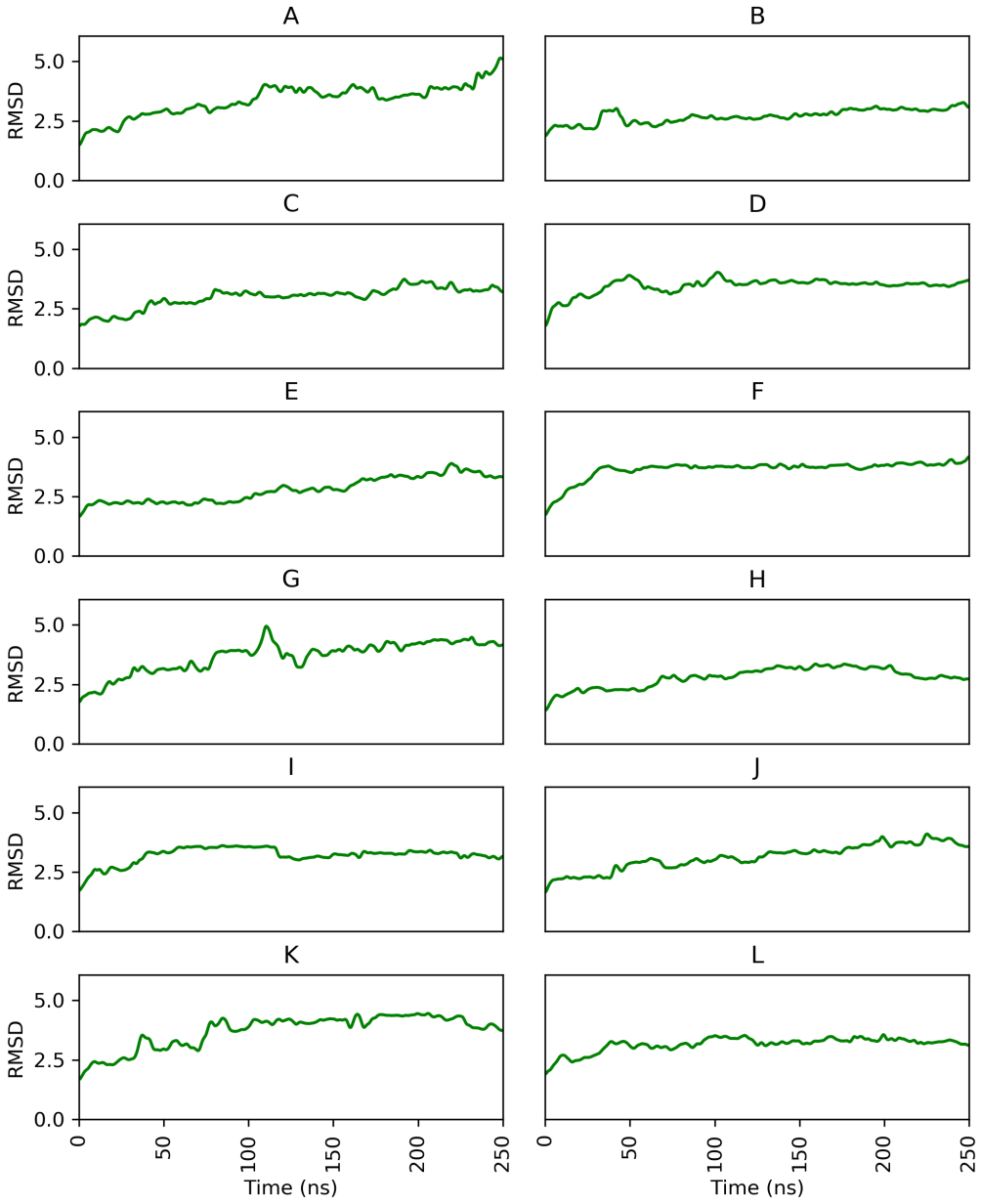


FIGURE C.1. RMSD over time for each system.

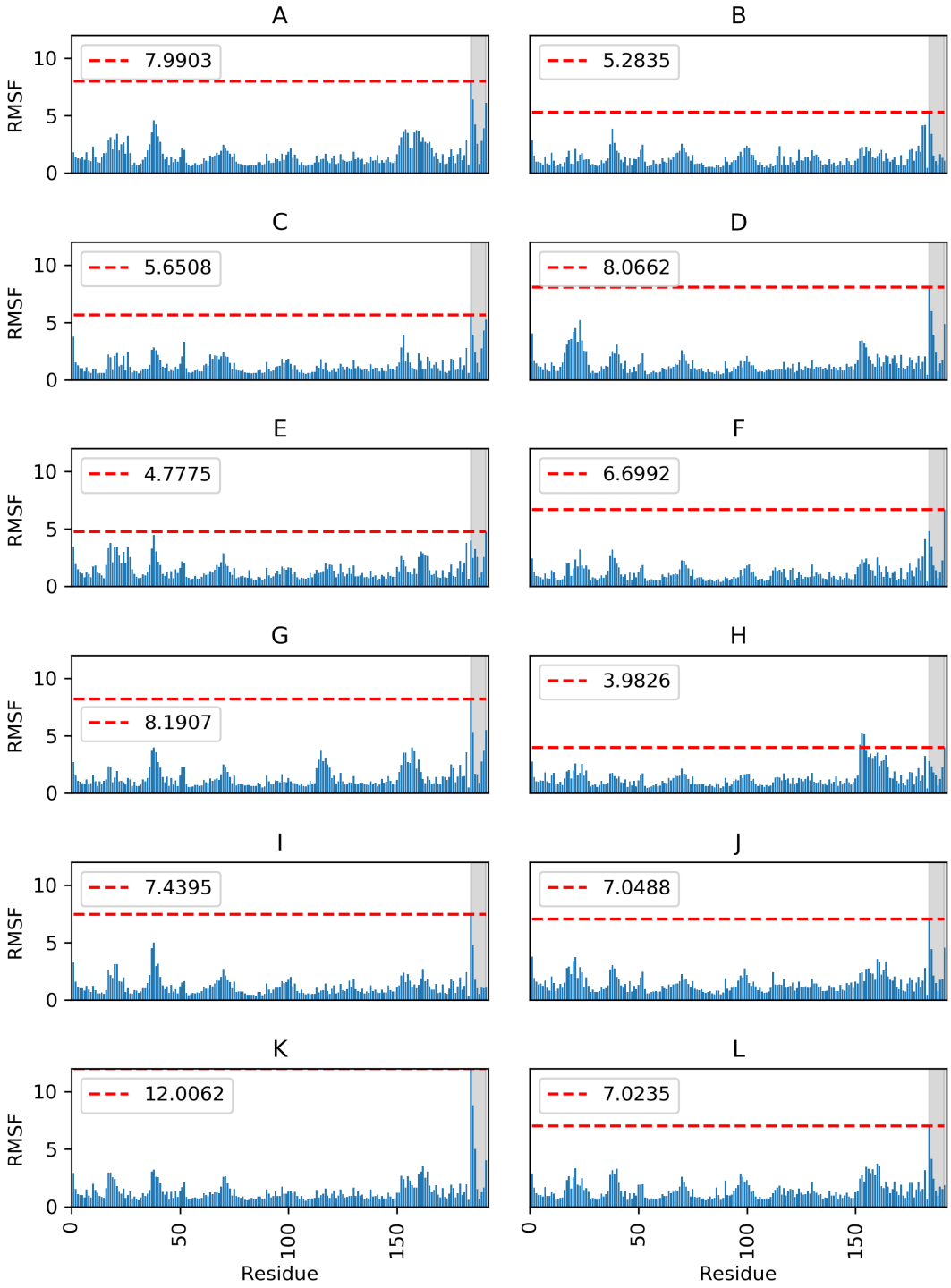


FIGURE C.2. RMSF for each system. Shaded region is the DNA substrate. Dashed red line indicates highest fluctuation value in the substrate nucleotides.

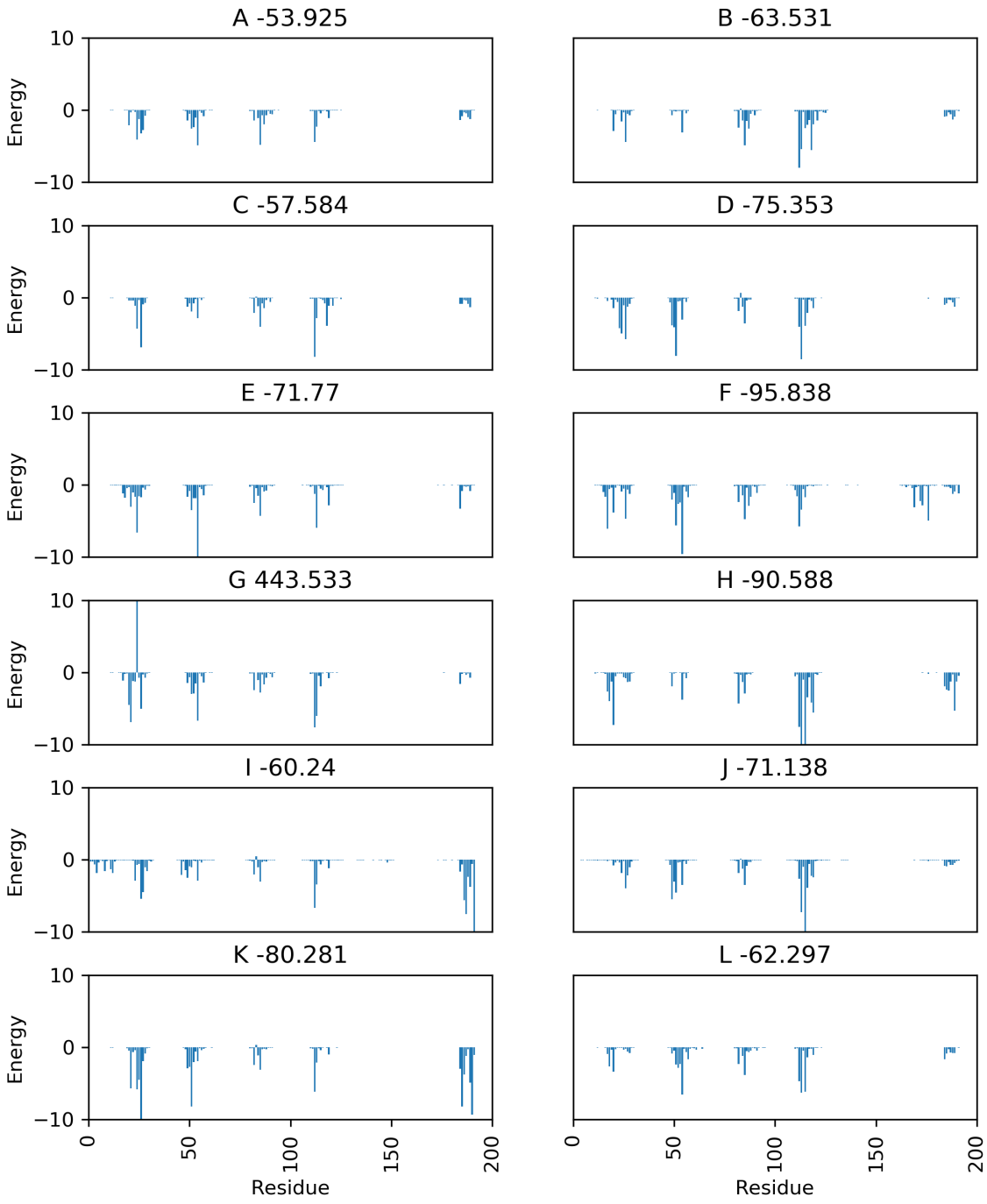


FIGURE C.3. Van der Waals interactions between each residue and the full DNA substrate. Each plot title includes the sum of the protein-substrate VdW interactions.

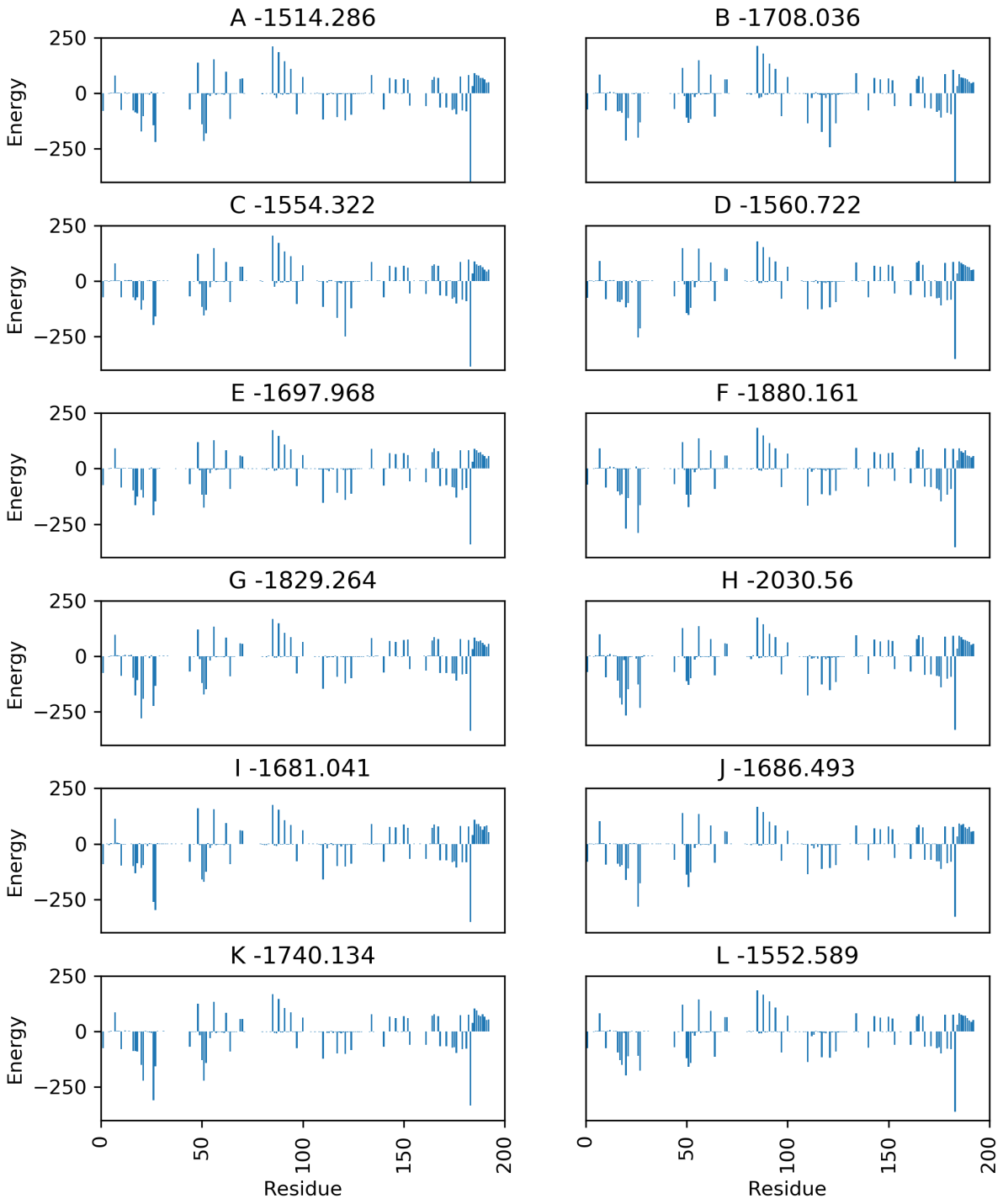


FIGURE C.4. Coulomb interactions between each residue and the full DNA substrate. Each plot title includes the sum of the protein-substrate Coulomb interactions.

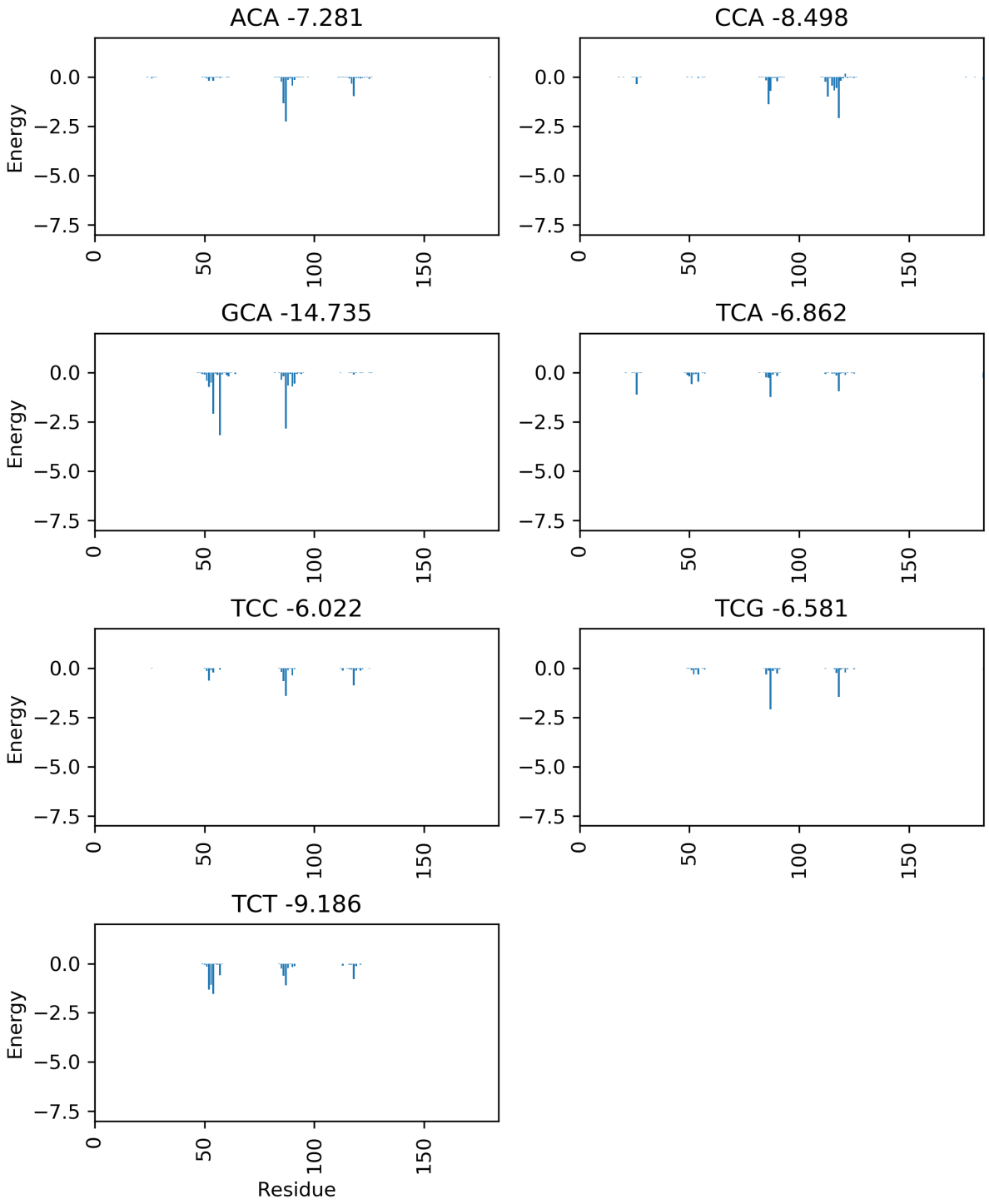


FIGURE C.5. Van der Waals interactions between each protein residue and 3' flanking nucleotide of target cytidine. Plot title indicates total of protein-nucleotide interactions

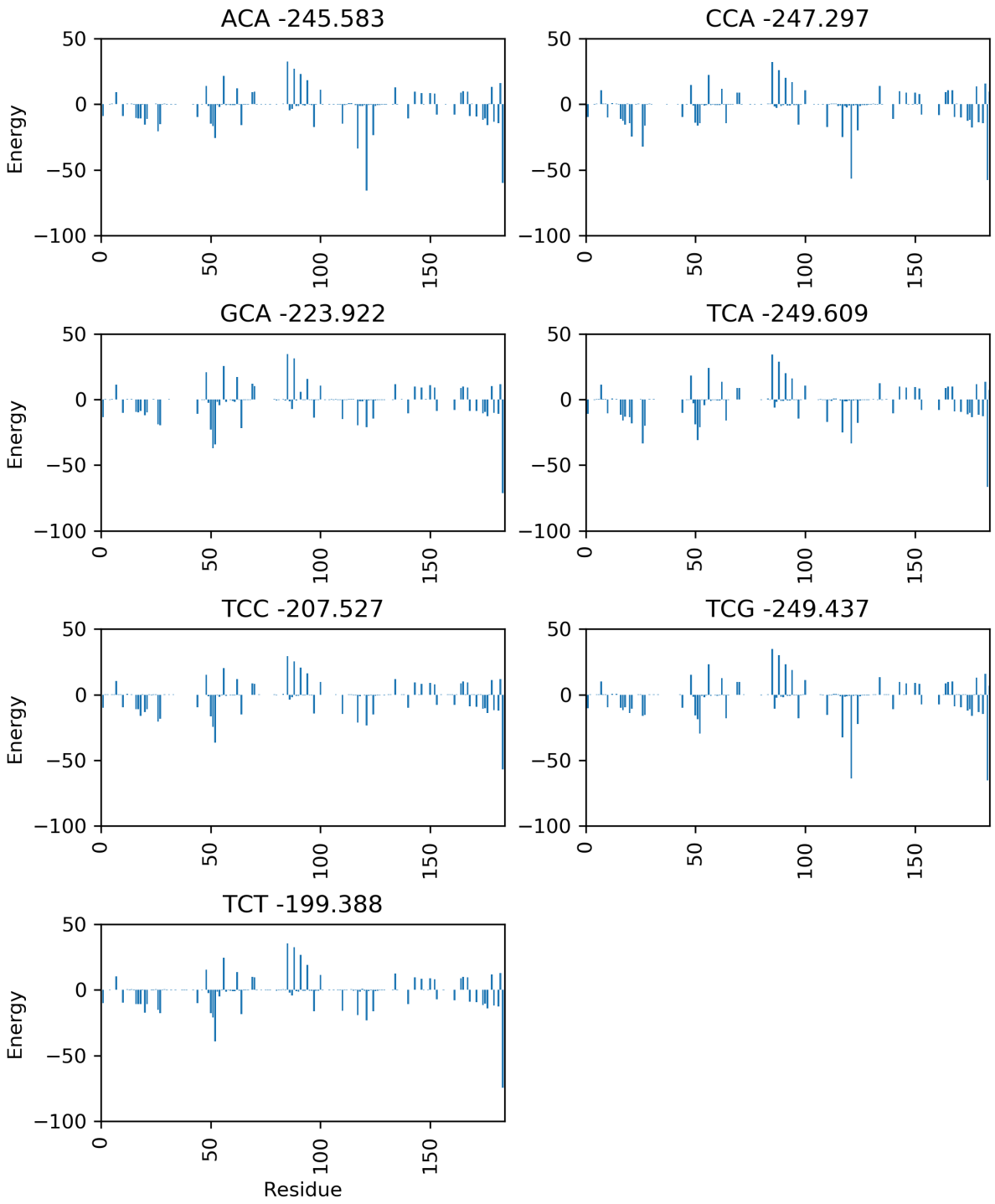


FIGURE C.6. Coulomb interactions between each protein residue and 3' flanking nucleotide of target cytidine. Plot title indicates total of protein-nucleotide interactions

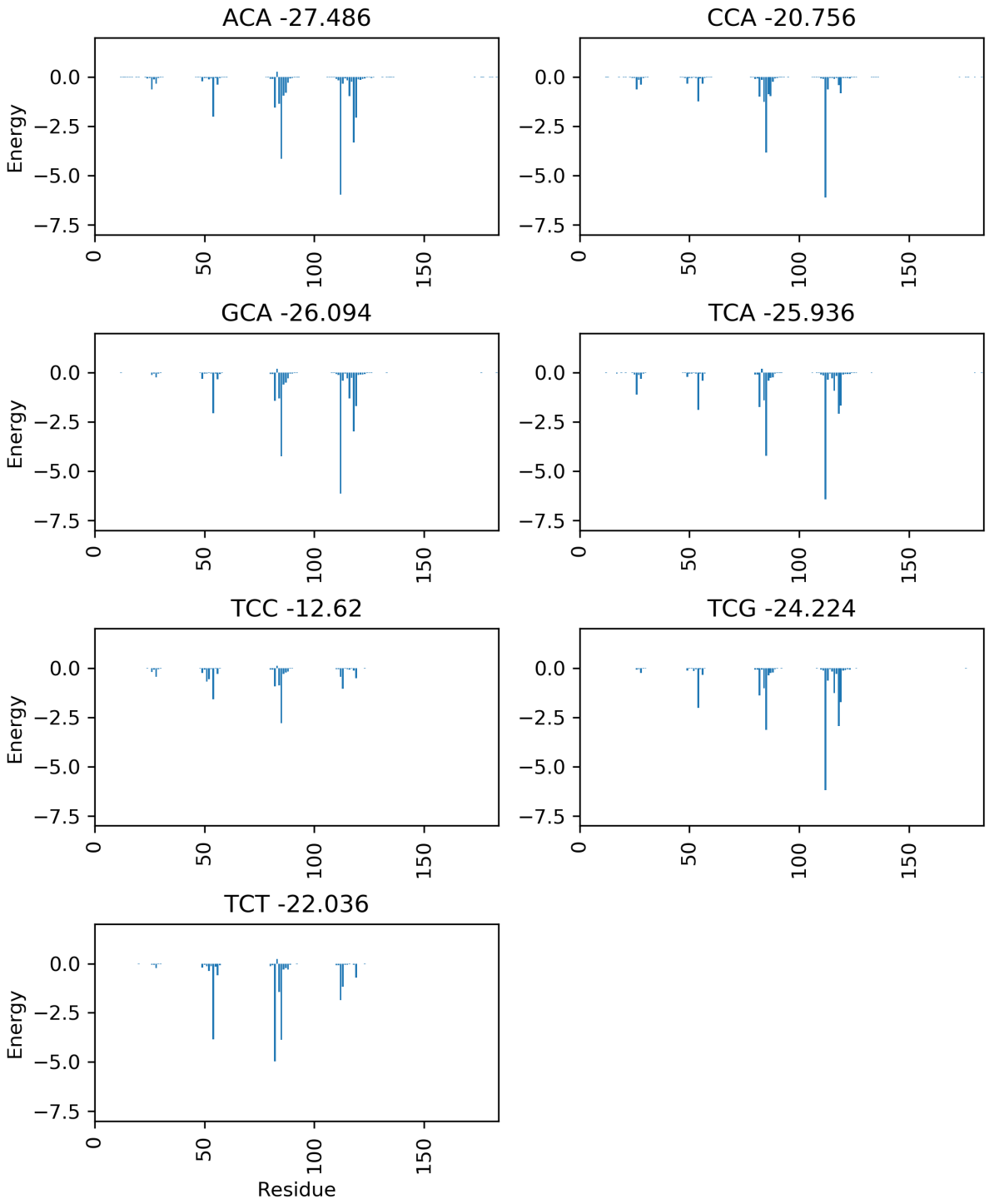


FIGURE C.7. Van der Waals interactions between each protein residue and target cytidine. Plot title indicates total of protein-nucleotide interactions

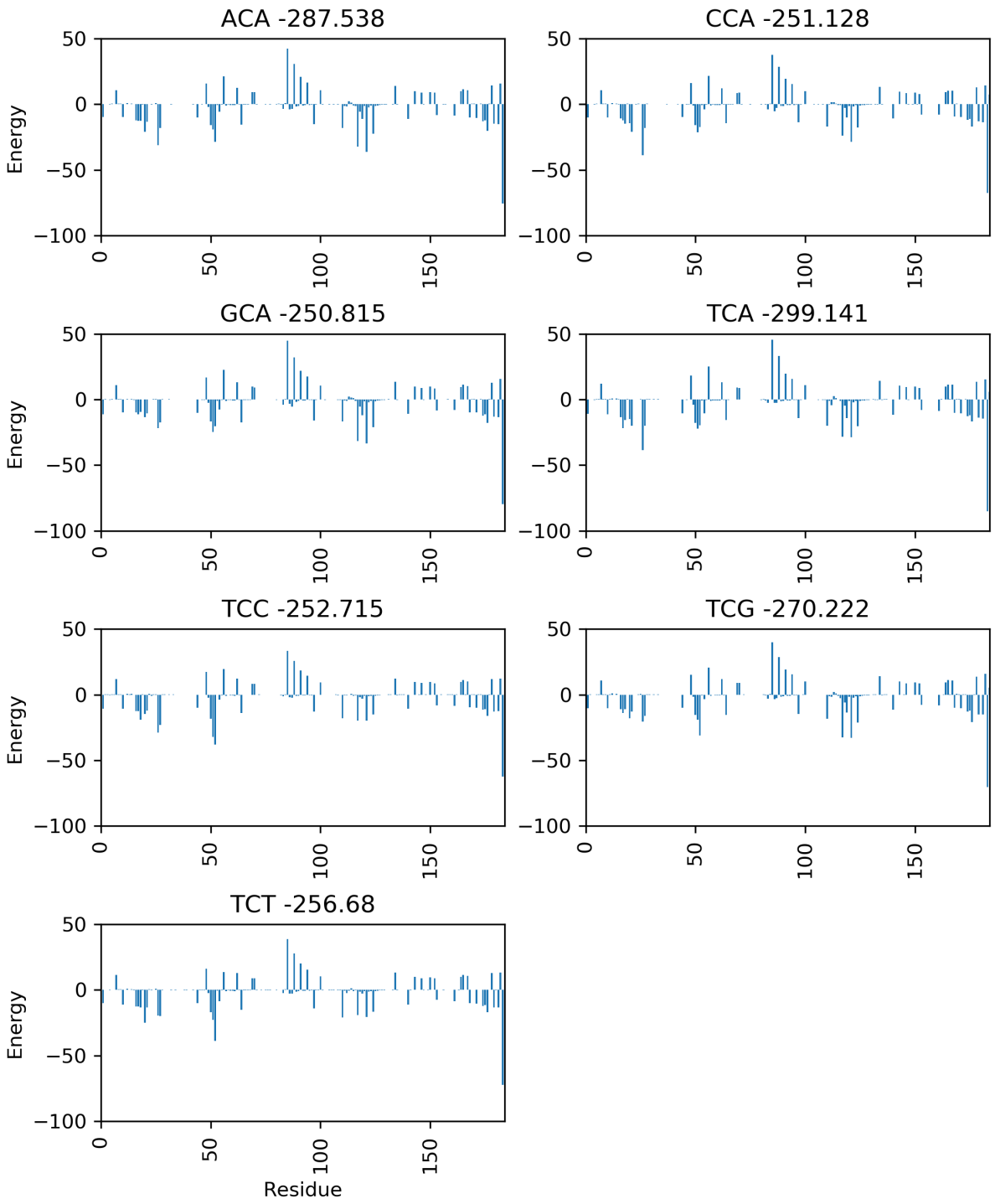


FIGURE C.8. Coulomb interactions between each protein residue and target cytidine. Plot title indicates total of protein-nucleotide interactions

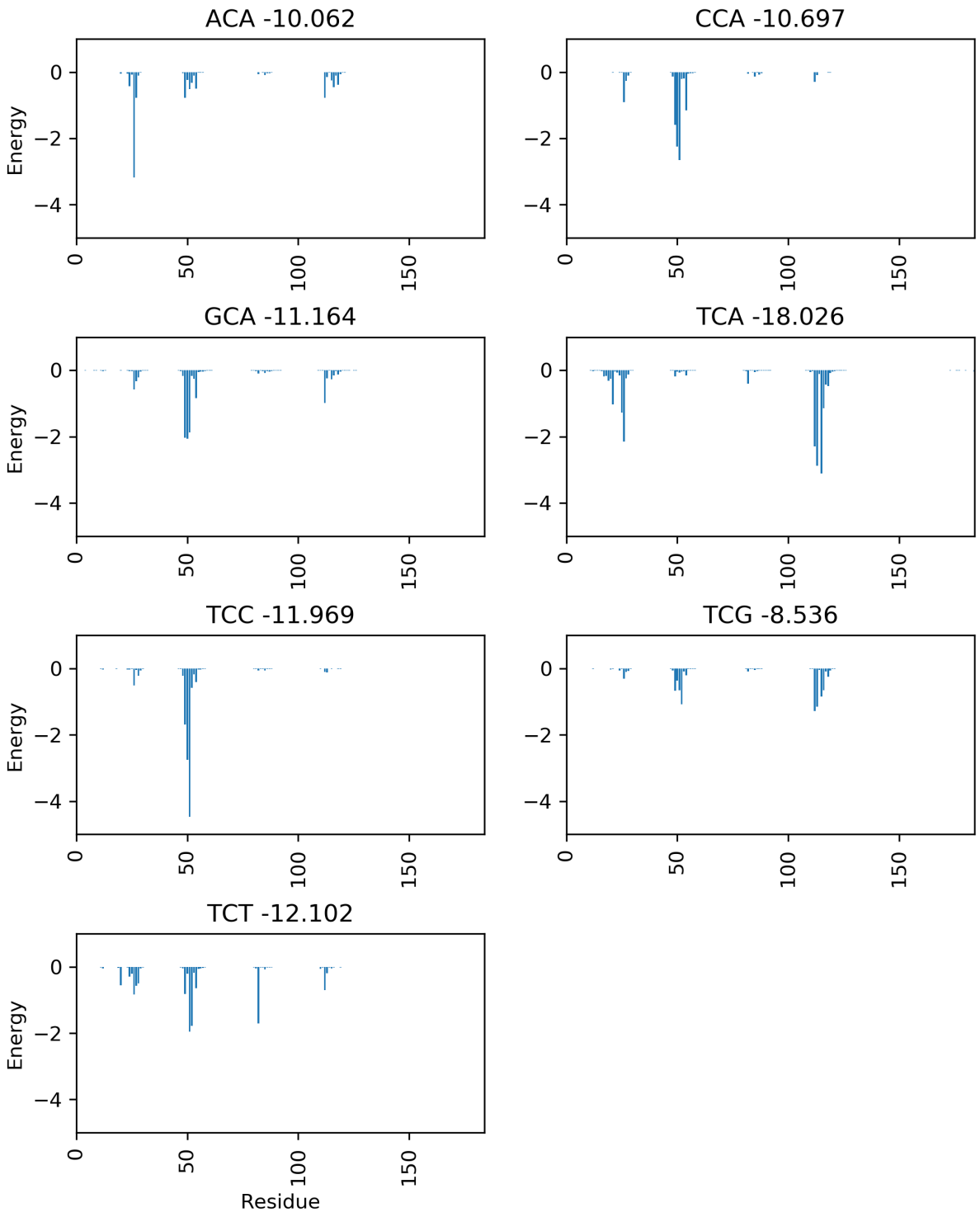


FIGURE C.9. Van der Waals interactions between each protein residue and 5' flanking nucleotide of target cytidine. Plot title indicates total of protein-nucleotide interactions

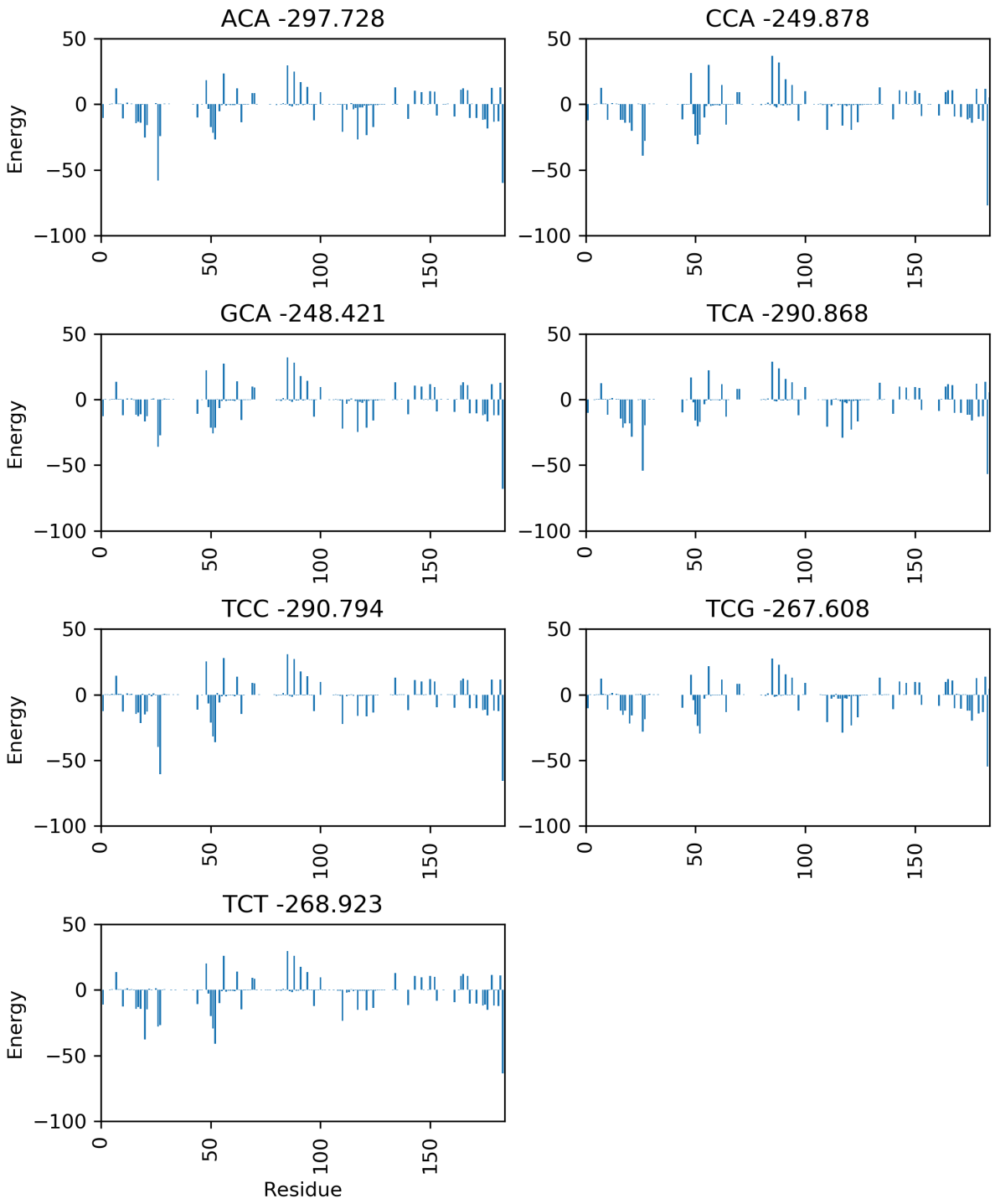


FIGURE C.10. Van der Waals interactions between each protein residue and 5' flanking nucleotide of target cytidine. Plot title indicates total of protein-nucleotide interactions

APPENDIX D

SINGLE-NUCLEOTIDE POLYMORPHISM OF THE DNA CYTOSINE DEAMINASE APOBEC3H HAPLOTYPE I LEADS TO ENZYME DESTABILIZATION AND CORRELATES WITH LUNG CANCER

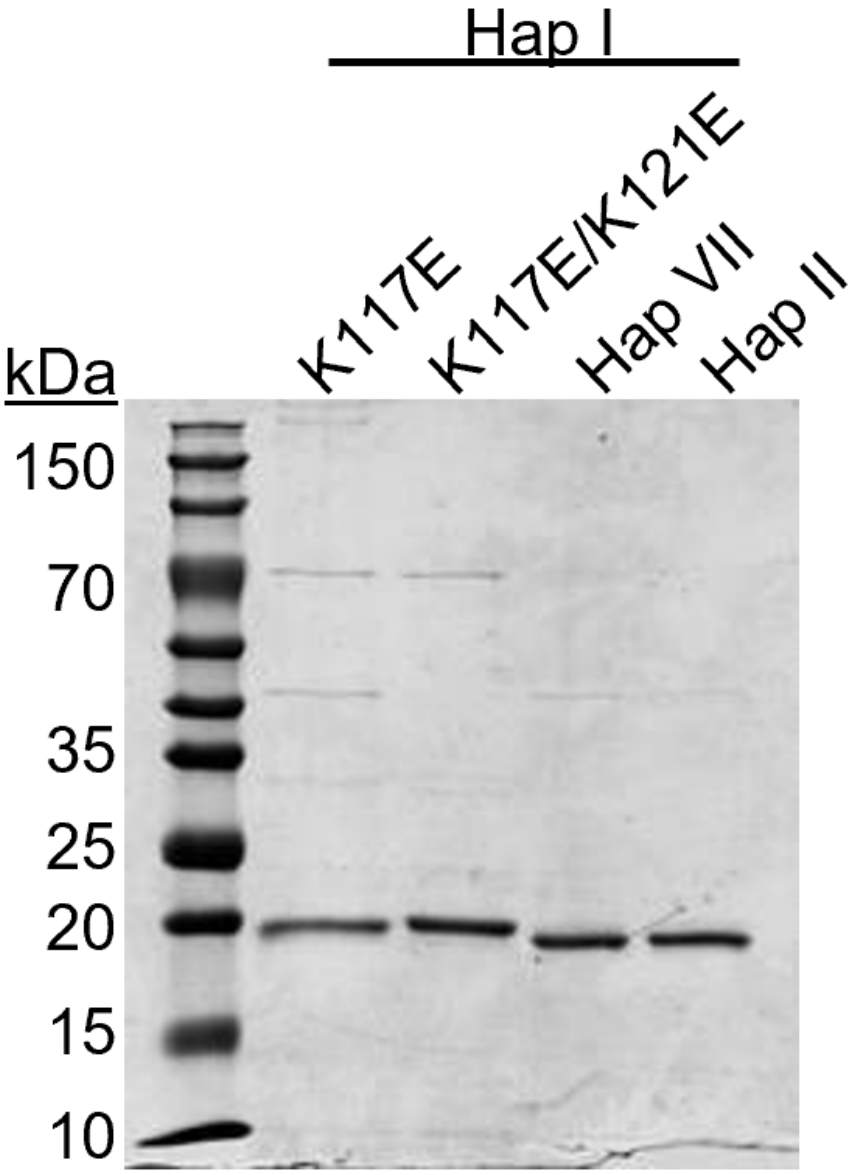


FIGURE D.1. **Purity of A3H enzymes.** Purified A3H enzymes were resolved by SDS-PAGE.

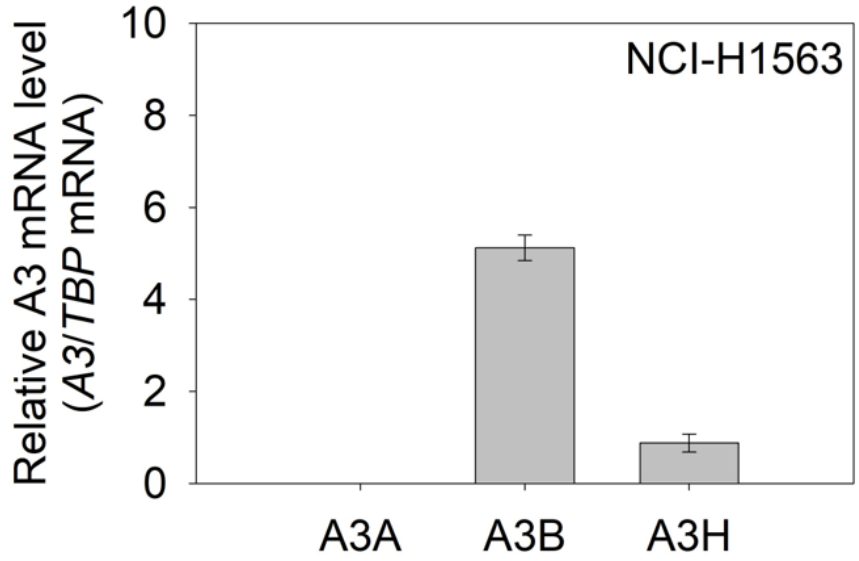


FIGURE D.2. **Detection of A3 mRNA by qPCR.** The endogenous A3A, A3B, and A3H mRNA levels were quantified in NCI-H1563 cells relative to TBP mRNA. The A3A mRNA was not detectable. Error bars represent the standard deviation from three independent experiments.

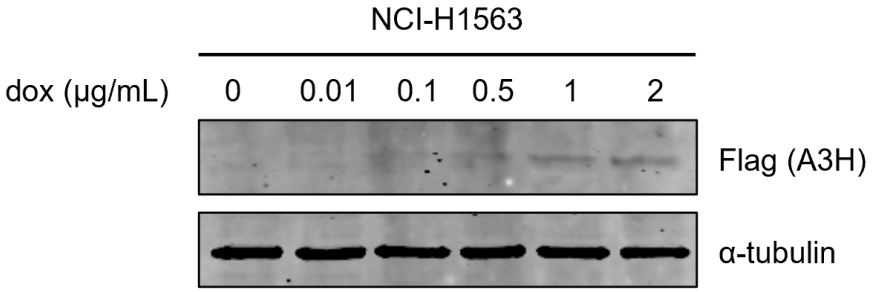


FIGURE D.3. **Expression of A3H Hap I-Flag in doxycycline (dox) inducible cell lines.** Titration of dox in NCI-H1563 cell lines shows that expression occurs at 1 µg/mL Dox or higher.

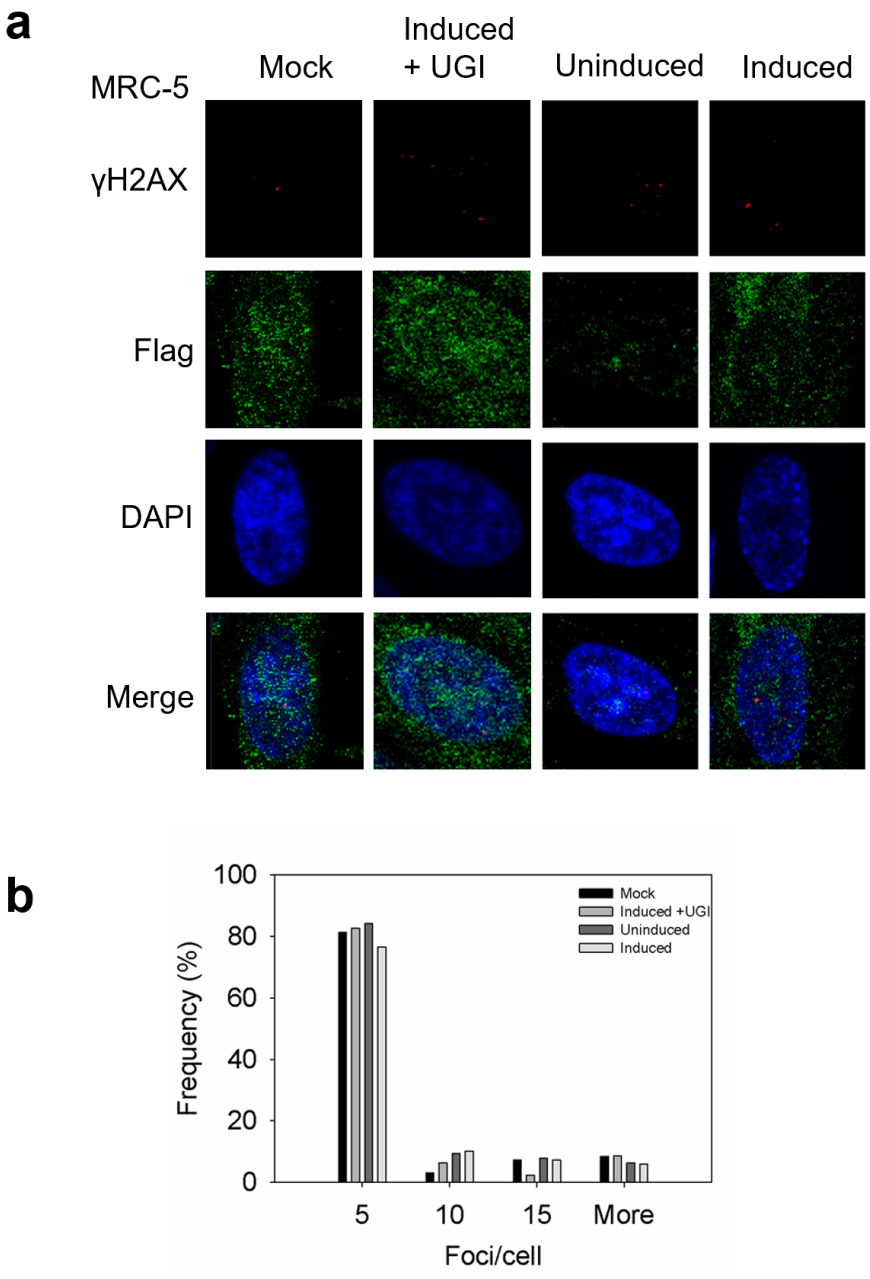


FIGURE D.4. **DNA damage induced by A3H Hap I.**

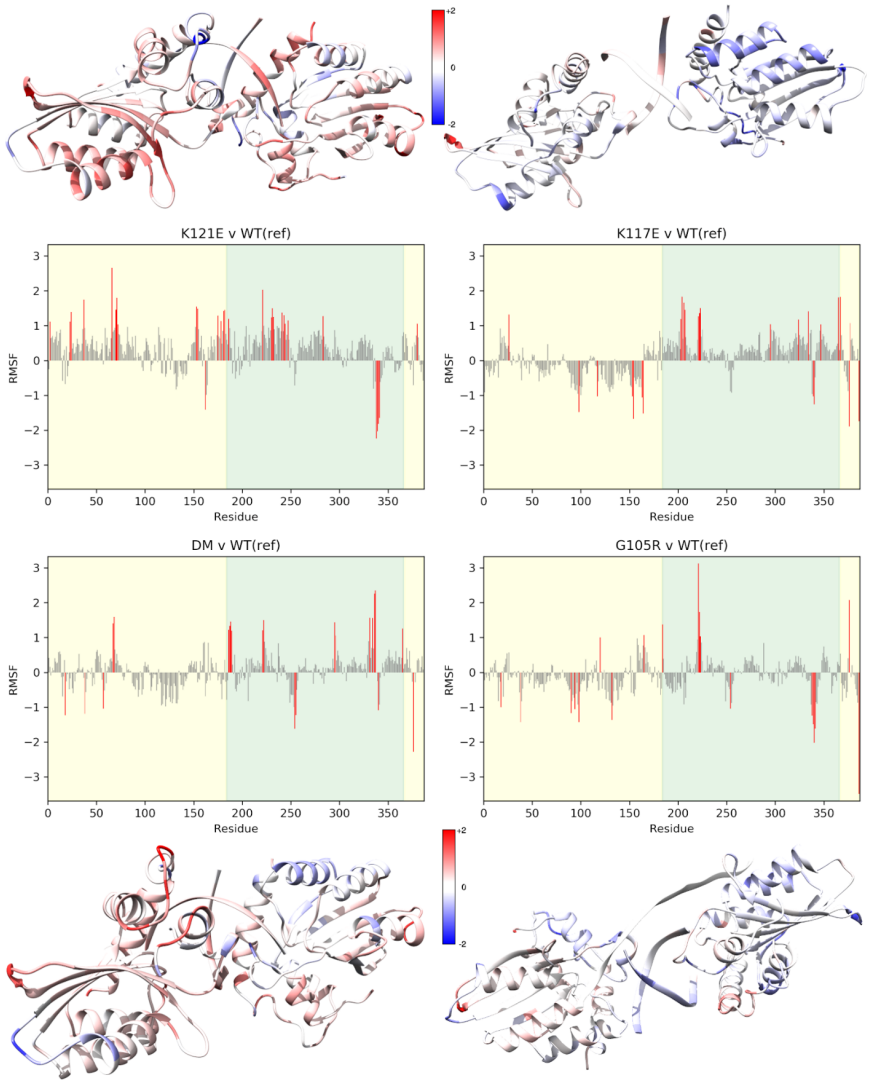


FIGURE D.5. **Differences in RMSF by residue for all systems compared to the A3H Hap I WT.** Residues highlighted in red change by more than 1Å.

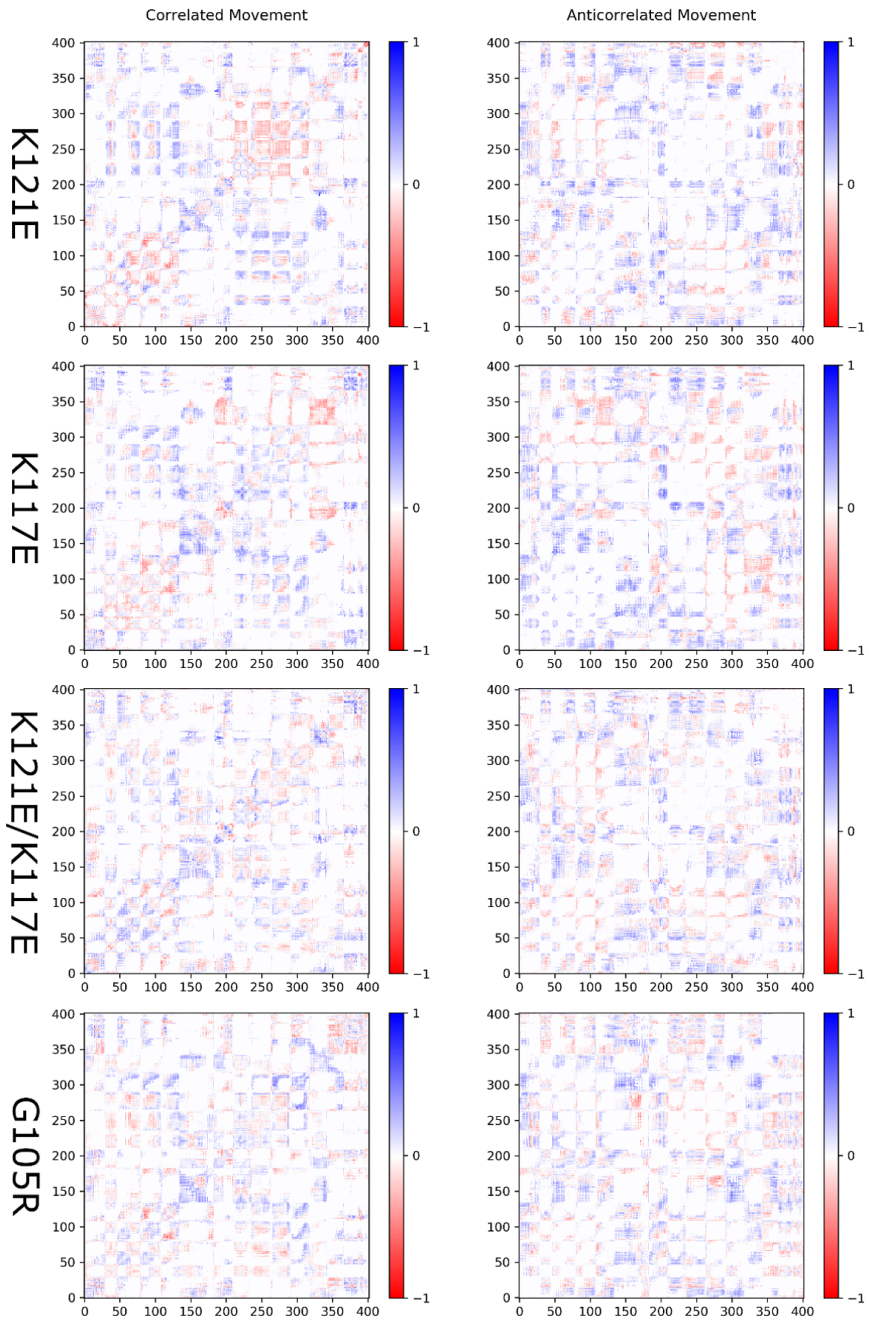


FIGURE D.6. **Difference correlation matrices comparing mutant systems to A3H Hap I WT.** Regions of blue exhibit an increased magnitude of correlation (left) or anticorrelation (right), regions of red indicate decreased magnitude.

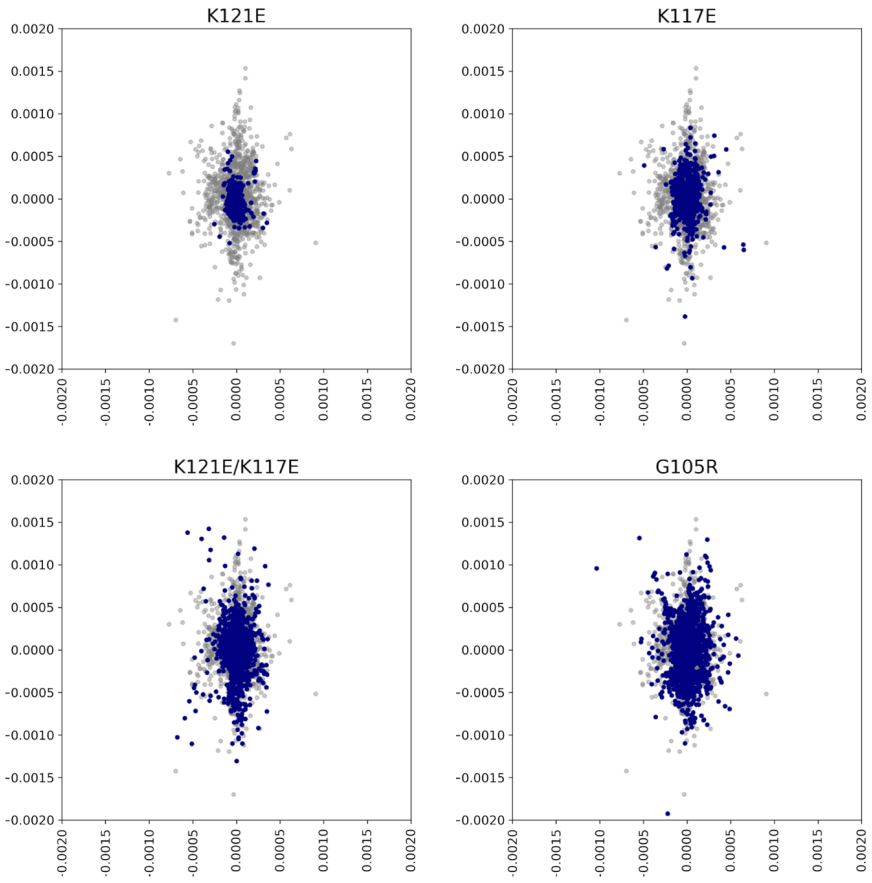


FIGURE D.7. **Principle Component Analysis (PCA) of all A3H systems comparing the first two modes of each system.** Points in grey are the A3H Hap I WT, which is shown as a reference for the individual mutations. K121E exhibits greatly diminished scatter in the first two modes. K117E shows slightly diminished scatter. K121E/K117E rescue and G105R both exhibit similar modes of motion to the wild type.

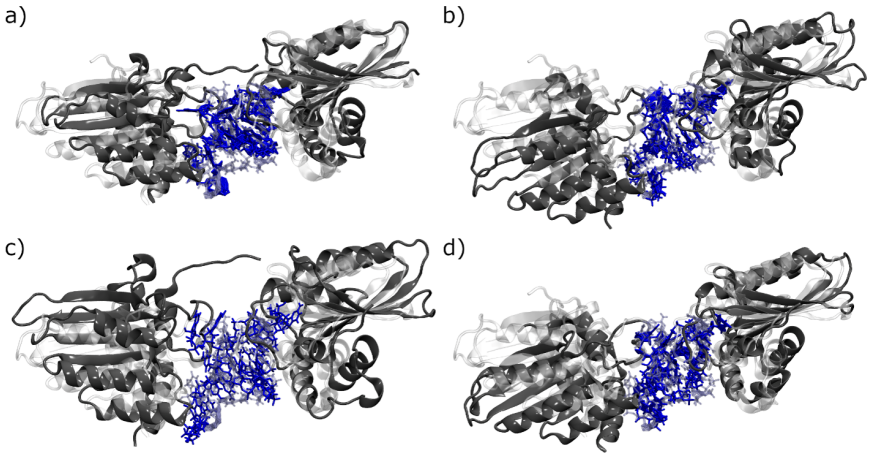


FIGURE D.8. **Average structure overlay.** The A3H Hap I WT is shown as transparency for **a)** K121E, **b)** K117E, **c)** K121E/K117E, and **d)** G105R.

	K+ ions added to neutralize
WT	12
K121E	16
K117E	16
K121E/K117E	20
G105R	10

TABLE D.1. Potassium Counterions added to each A3H dimer system to bring total system charge to 0.

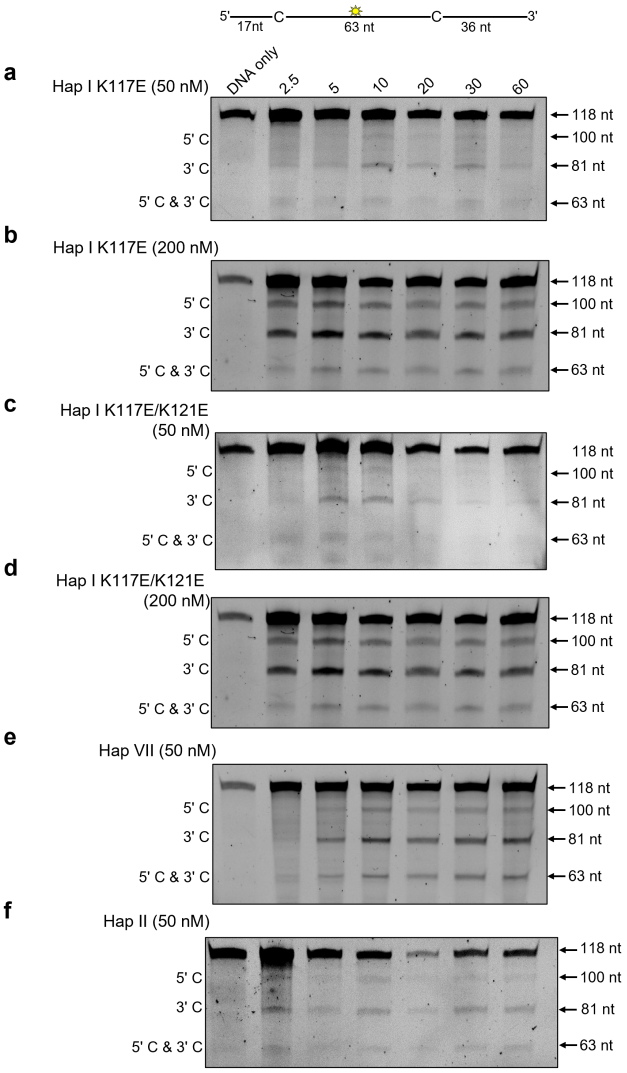


FIGURE D.9. **Deamination activity of A3H Haplotypes and mutants.** Time course of deamination of A3H Hap I mutants, A3H Hap II, and A3H Hap VII at 50 nM or 200 nM of enzyme. Deamination was tested on a 118 nt ssDNA (100 nM) and gels analyzed for each plot were quantified and plotted as shown in Figure 5.4b.

	WT	K121E	K117E	K121E/K117E	G105R
K/E121 - K/E117	56.61%	65.39%	55.36%	53.01%	0.76%
K/E121 - R124	0.77%	38.65%	0.31%	41.65%	0.09%
R124 - I182	2.93%	19.26%	2.16%	6.24%	0.68%
P118-S86	4.46%	47.28%	33.70%	26.26%	0.00%
R175 - G374	35.27%	4.39%	5.19%	44.88%	39.77%
R175 - G375	24.78%	6.20%	33.75%	35.76%	3.23%
R175 - C376	0.00%	10.25%	0.00%	0.00%	0.68%
R176 - G374	0.06%	30.63%	9.15%	0.01%	4.86%
R176 - G375	27.23%	44.87%	76.64%	76.47%	22.92%
R176 - C376	10.10%	10.06%	4.85%	87.29%	2.77%
R179 - G374	24.88%	0.03%	0.00%	0.00%	7.21%
R179 - G375	27.72%	23.36%	10.06%	60.68%	30.01%
R179 - C376	1.71%	52.30%	43.21%	84.41%	9.69%
Q120 - C376	41.37%	4.65%	14.17%	86.92%	0.24%

TABLE D.2. Largest contributors to hydrogen bonding interactions presented as percentage of total simulation time. These interactions form a network across a portion of the protein surface including active site residues and RNA-binding residues.

APPENDIX E

DIVERGENCE IN DIMERIZATION AND ACTIVITY OF PRIMATE APOBEC3C

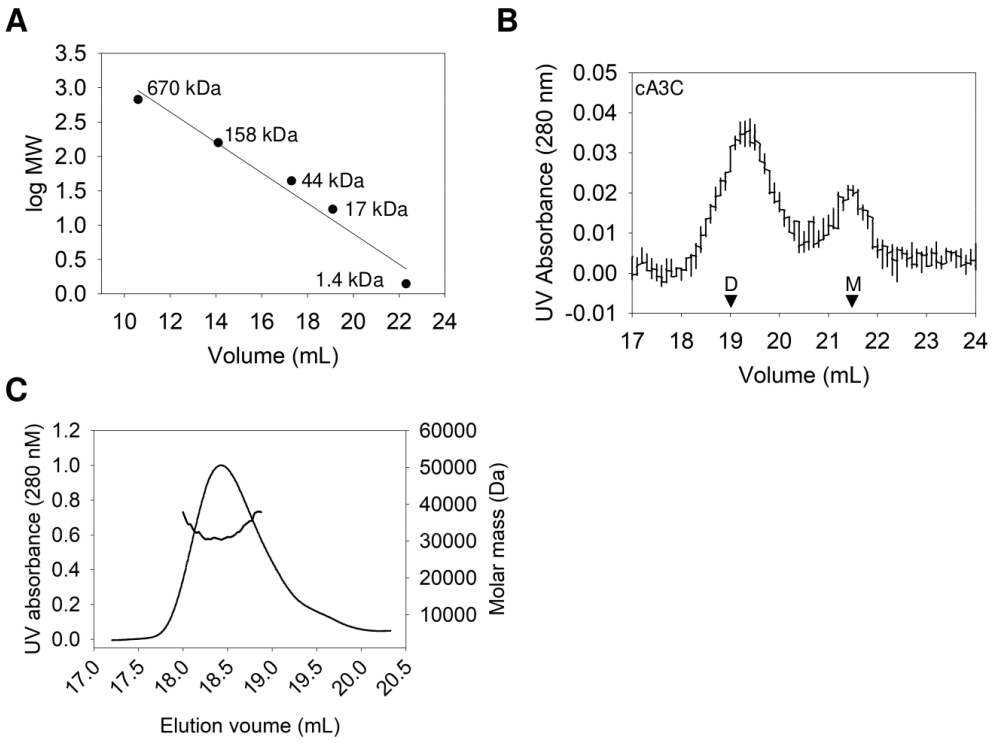


FIGURE E.1. **Standards used for size exclusion chromatography. (A)** The standard curve generated from the G200 Increase column. Sizes correspond to Thyroglobulin (670 kDa), Gamma globulin (158 kDa), Ovalbumin (44 kDa), Myoglobin (17 kDa), and Vitamin B12 (1.4 kDa). **(B)** The cA3C is a mix of dimer (D) and monomer (M). The cA3C SEC was carried out as for rhA3C and showed that monomers elute at a volume greater than 20 mL and dimers elute at a volume smaller than 20 mL. **(C)** To confirm our measurements by SEC, which is based on standards, we also conducted multiangle light scattering (MALS) on the rhA3C R44Q/R45H/N115K, which had a single peak. We empirically calculated a molar mass of 30 kDa at the peak of elution, which is near the predicted molar mass from the amino acid sequence of 23 kDa.

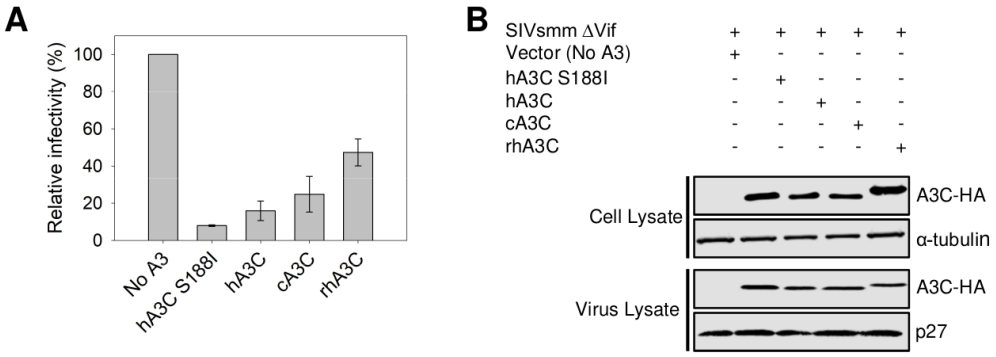


FIGURE E.2. **A3C-mediated restriction of SIVsmm. (A)** SIVsmm L5 ΔVif infectivity was measured by β- galactosidase expression driven by the HIV-1 5 LTR from TZM-bl cells infected with SIVsmm that was produced in the absence or presence of 3xHA tagged hA3C, hA3C S188I, cA3C, and rhA3C. Results normalized to the no A3 condition are shown with the Standard Deviation of the mean calculated from at least three independent experiments. **(B)** Immunoblotting for the HA tag was used to detect A3C enzymes expressed in cells and encapsidated into SIVsmm ΔVif virions. The cell lysate and virion loading controls were α-tubulin and p27, respectively.

Mutations	ΔE compared to wild type (kcal mol^{-1})
R44Q	+161.6
R45H	+384.4
N115M	+173.8
A144S	-13.0
R44Q/R45H	+643.3
R44Q/R45H/N115M	+631.3
R44Q/R45H/A144S	+769.8

TABLE E.1. **Dimer Interaction Energy Decomposition Analysis.** Values shown as comparisons to wild type interaction energy.

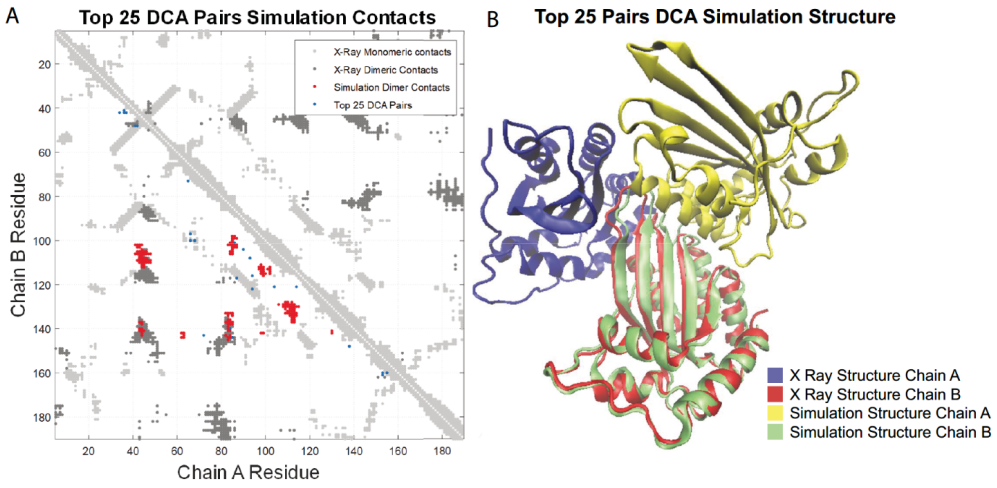


FIGURE E.3. **Structure based model molecular dynamics simulation driven by coevolved pairs is used to predict dimeric complexes.** A simulation utilizing the top dimeric residue-residue couplings estimates a putative dimeric complex. **(A)** shows a comparison of the dimeric interactions found in the 3VOW crystal structure (dark grey dots) and the dimeric interactions identified by the homodimer prediction (red dots). **(B)** Although the predicted complex and the crystal structure are not identical, they do share several important dimeric interactions, particularly those involving the residue 144.

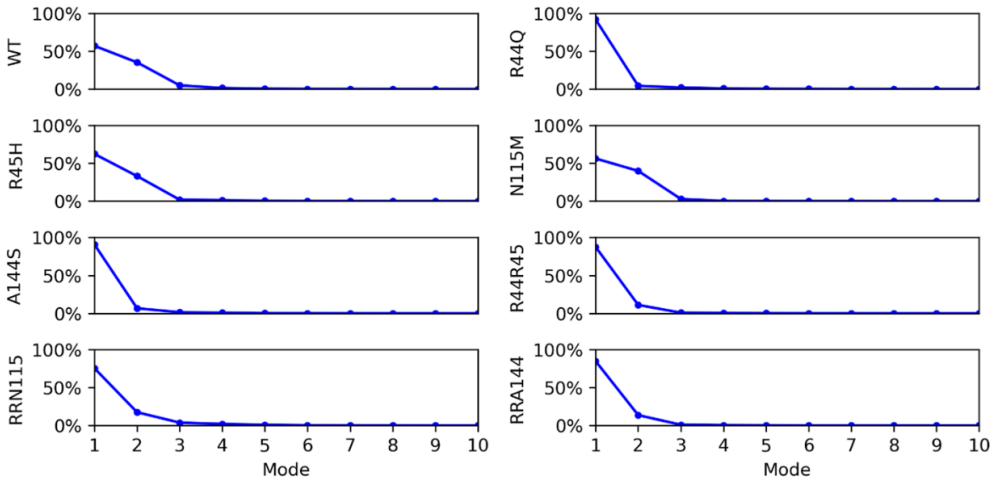


FIGURE E.4. **Contribution to total dynamic motion from first ten (10) normal modes.** Amino acid changes are listed on the y-axis. R44R45 is the combined mutant R44Q/R45H. RRN115 is the combined mutant R45H/R44Q/M115N. RRA144 is the combined mutant R45H/R44Q/S144A.

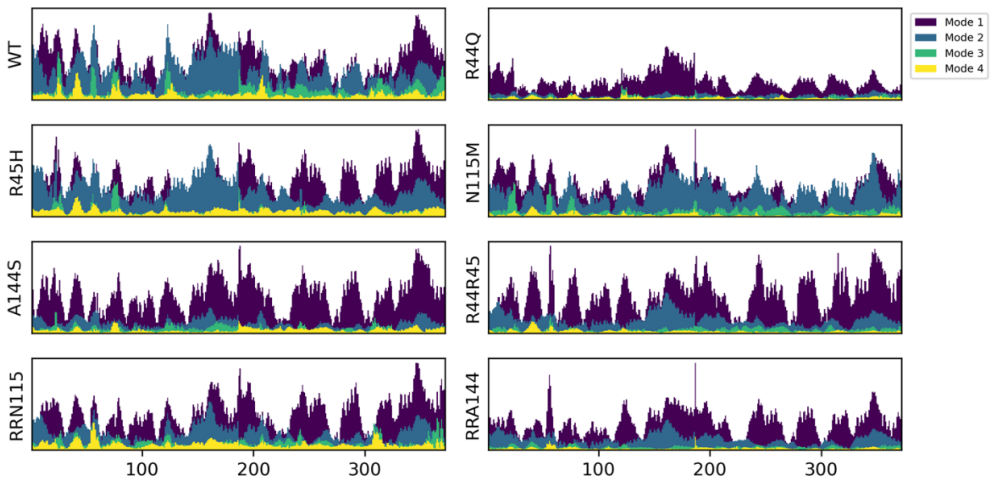


FIGURE E.5. **First four (4) normal modes by residue.** As shown above, most of the essential motion is captured by the first three modes. The fourth mode is included here to show that it is negligible. R44R45 is the combined mutant R44Q/R45H. RRN115 is the combined mutant R45H/R44Q/M115N. RRA144 is the combined mutant R45H/R44Q/S144A.

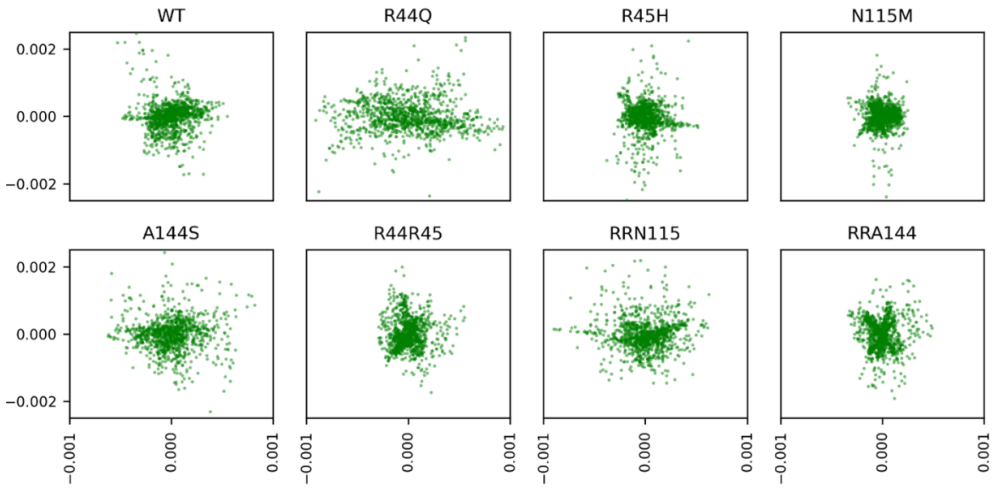


FIGURE E.6. **PCA for first two normal modes.** R44R45 is the combined mutant R44Q/R45H. RRN115 is the combined mutant R45H/R44Q/M115N. RRA144 is the combined mutant R45H/R44Q/S144A.

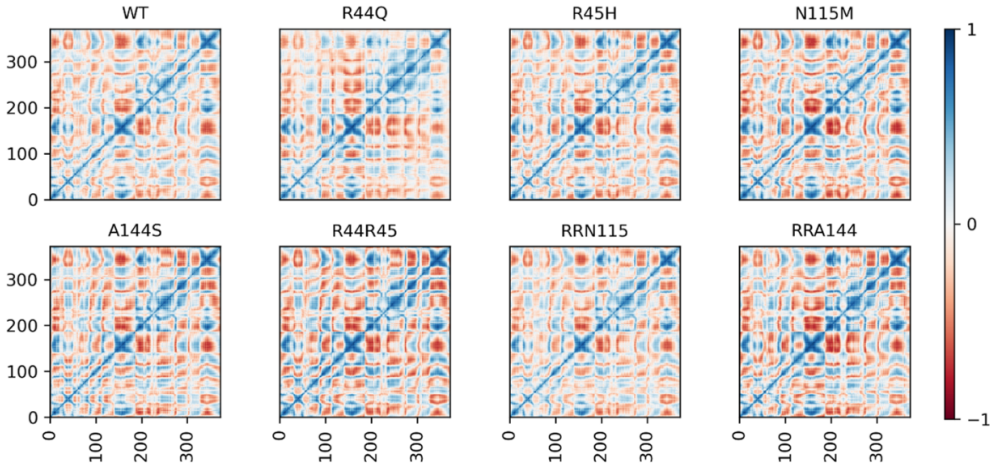


FIGURE E.7. **Correlated motion matrices pairwise by amino acid.** R44R45 is the combined mutant R44Q/R45H. RRN115 is the combined mutant R45H/R44Q/M115N. RRA144 is the combined mutant R45H/R44Q/S144A.

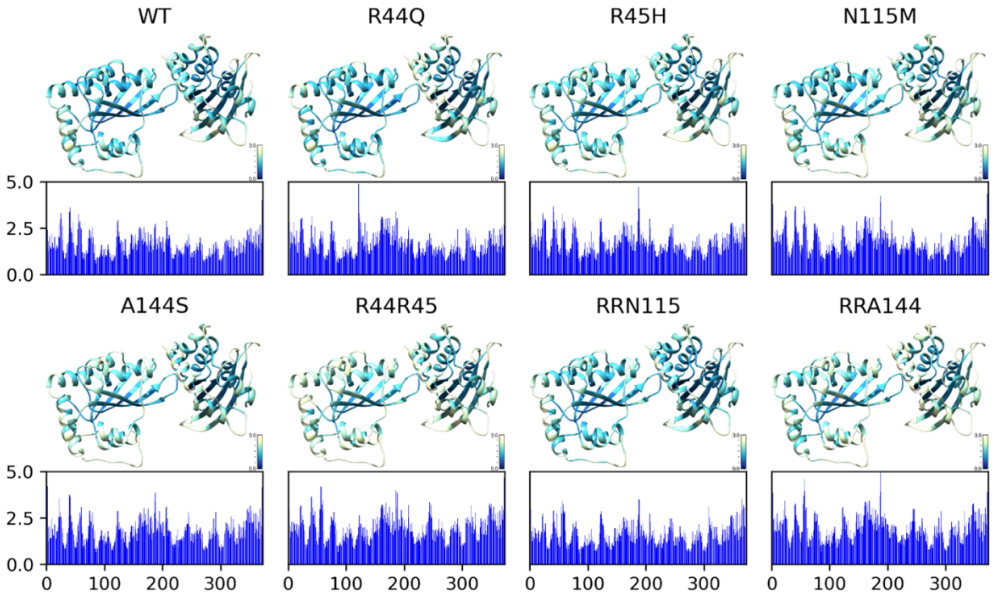


FIGURE E.8. **RMSF by residue mapped to 3D structure.** Barplots below structures show raw RMSF data. R44R45 is the combined mutant R44Q/R45H. RRN115 is the combined mutant R45H/R44Q/M115N. RRA144 is the combined mutant R45H/R44Q/S144A.

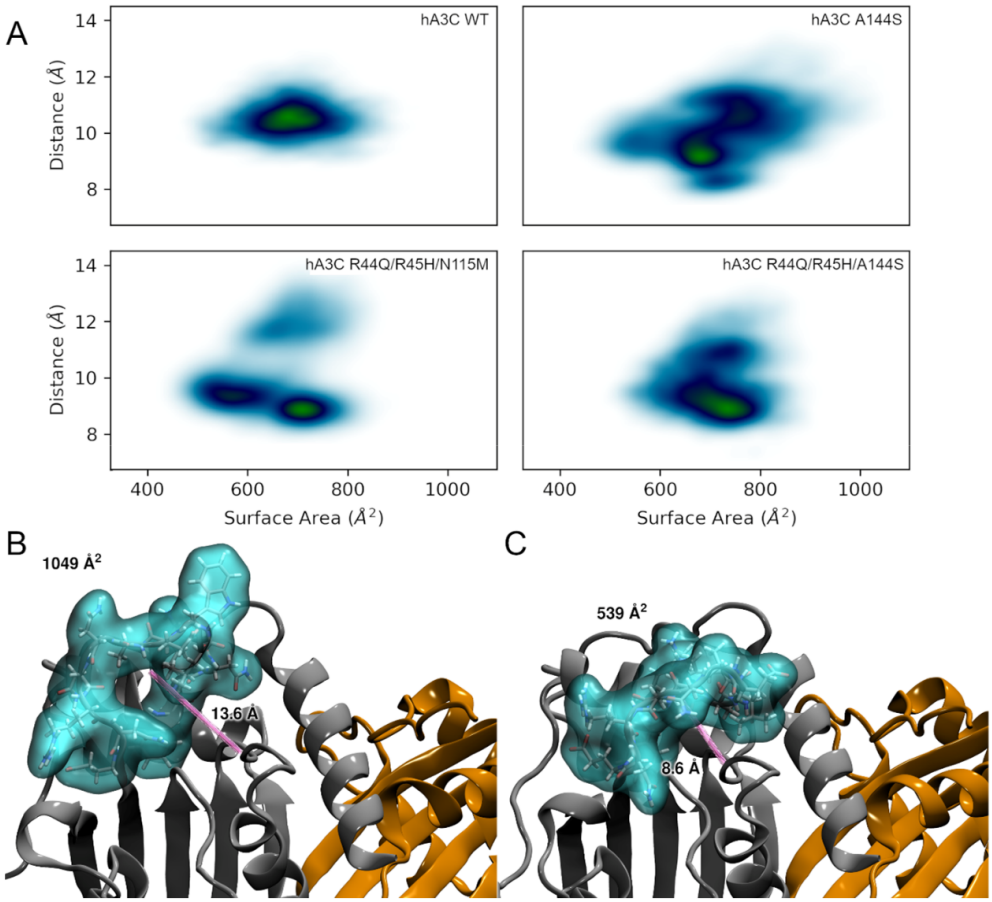


FIGURE E.9. **Analysis of loop 1.**

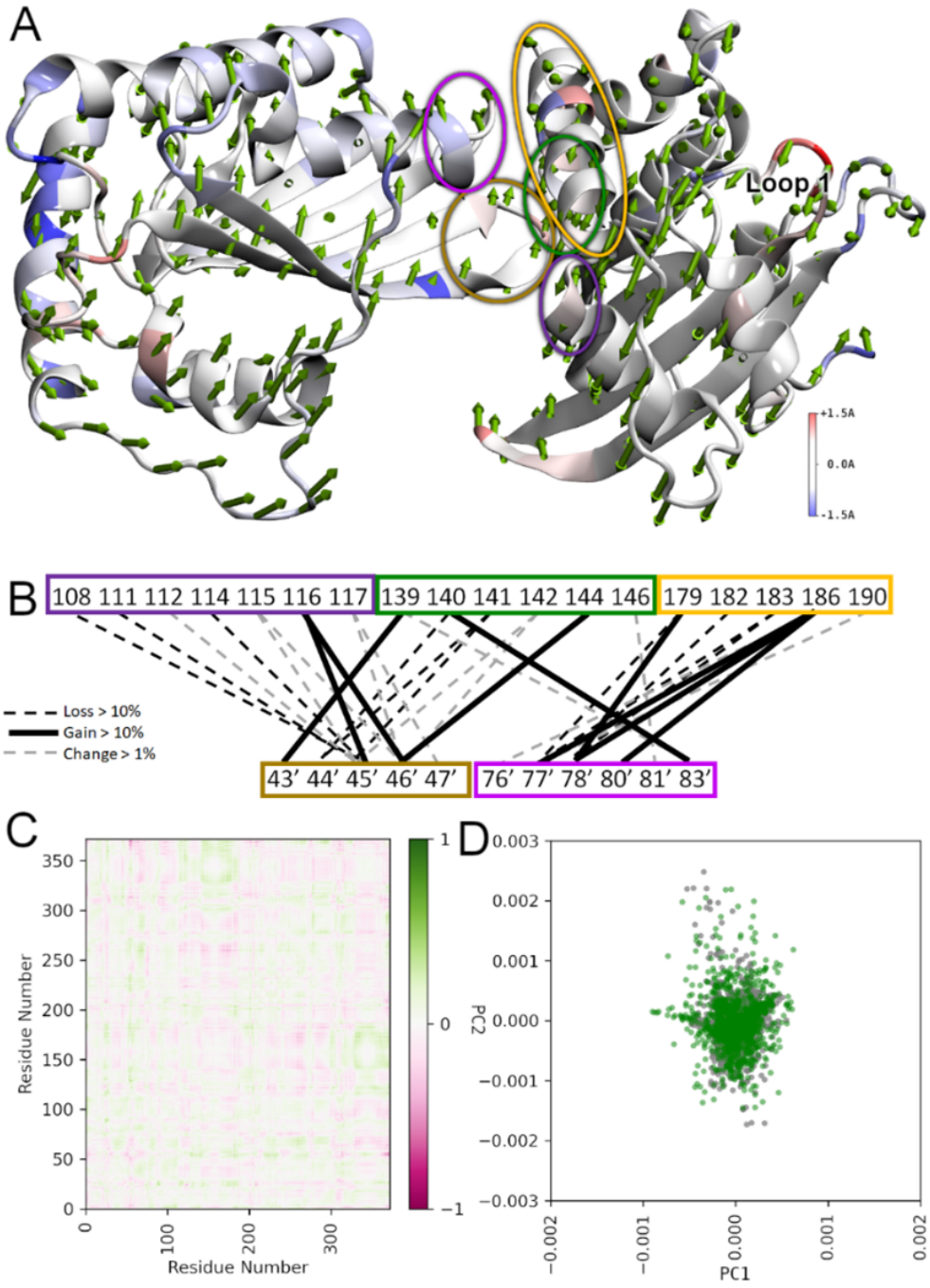


FIGURE E.10. **Model analysis of hA3C R44Q/R45H/N115M illustrates changes from hA3C WT.**

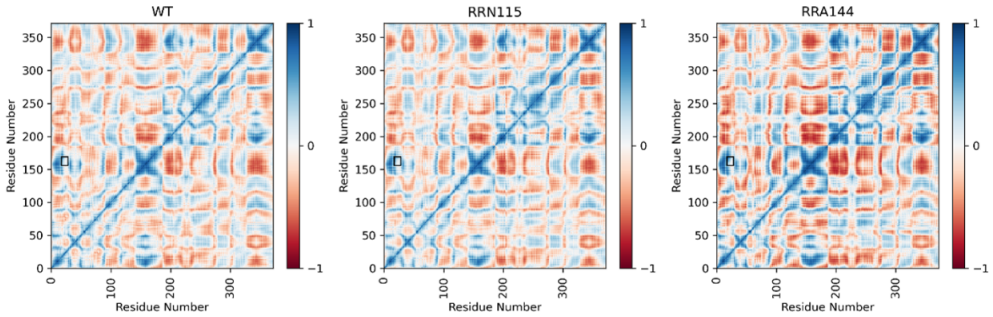


FIGURE E.11. **Correlated motion matrices pairwise by amino acid.**Correlated motion of hA3C WT, R44Q/R45H/N115M, and R44Q/R45H/A144S variants with loop 1/loop 7 correlated motion highlighted (black box). RRN115 is the combined mutant R44Q/R45H/N115M. RRA144 is the combined mutant R44Q/R45H/A144S.

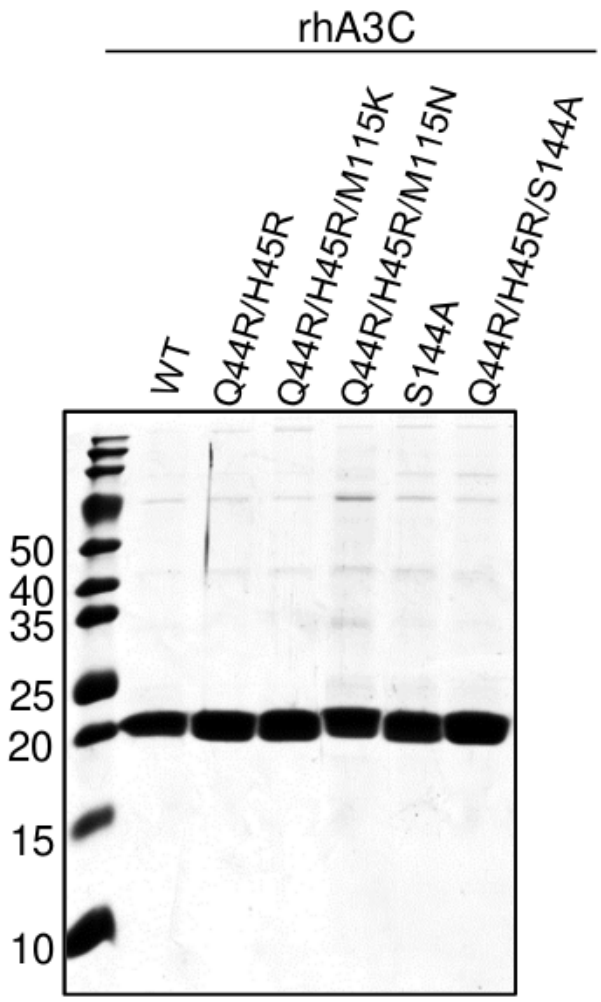


FIGURE E.12. **Purified rhA3C.** Five micrograms of rhA3C WT and mutants were resolved by SDS-PAGE.

APPENDIX F

INVESTIGATION OF DRIVERS FOR PREFERENTIAL CARBOXY-*S*-ADENOSYL-L-METHIONINE SYNTHESIS BY M.MPEI VARIANTS

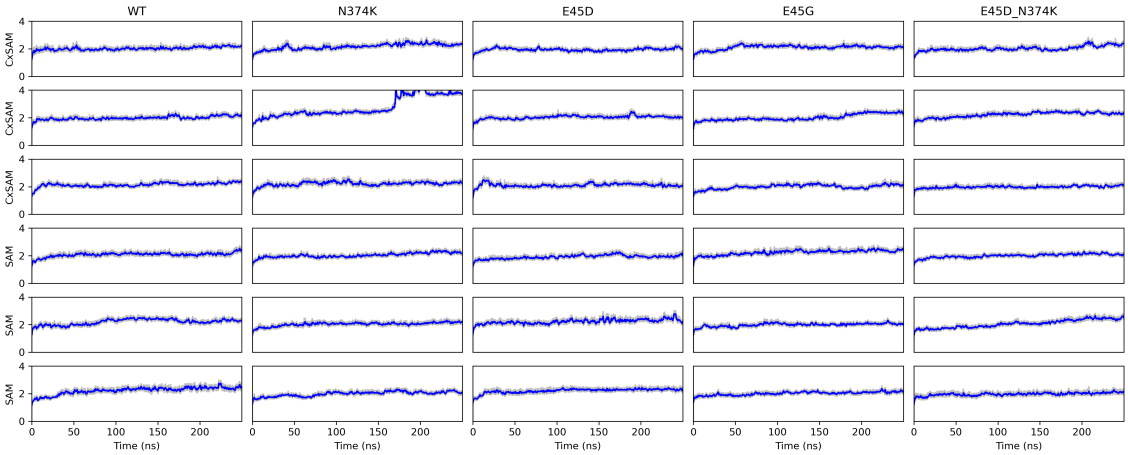


FIGURE F.1. Root-mean-squared deviation (RMSD) of all variant/cosubstrate combinations in triplicate. Systems are stable and show no large conformational shifts during dynamics.

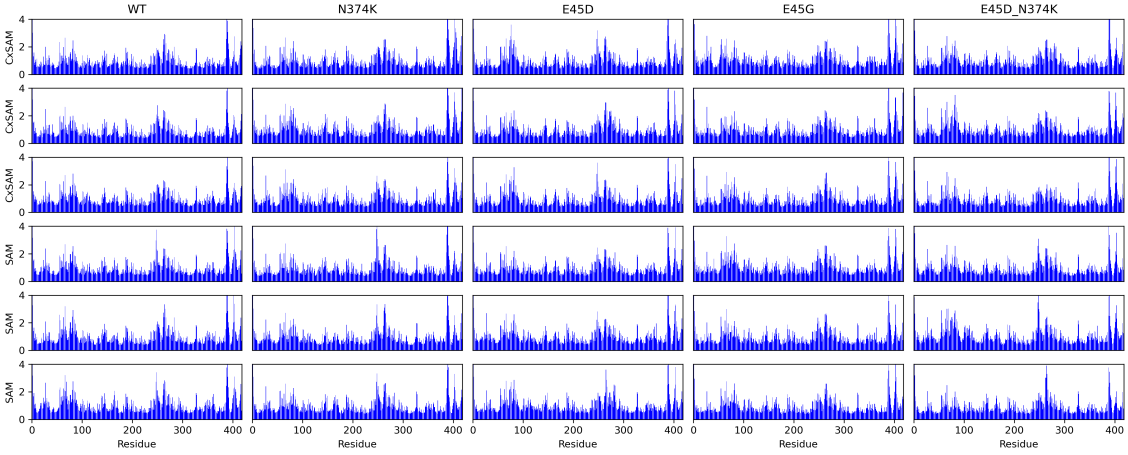


FIGURE F.2. Root-mean-squared fluctuation (RMSF) of all variant/cosubstrate combinations in triplicate. Each plot shows the fluctuation by residue of each system. The overall RMSF profile does not significantly change across systems, showing no significant differences in dynamic motion due to point mutations or changed cosubstrate.

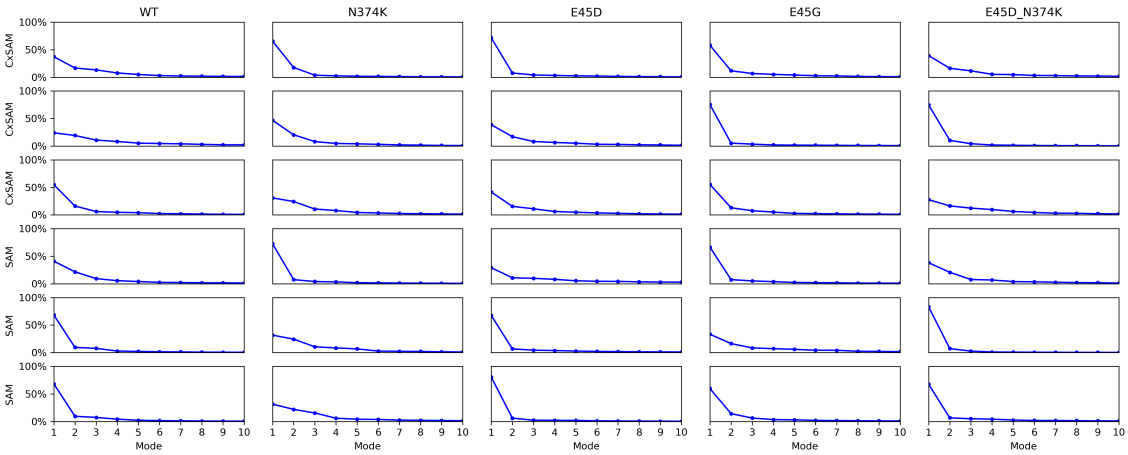


FIGURE F.3. Largest contributions to essential motion for each variant/cosubstrate combination in triplicate. In all cases, >95% of motion is captured by the first three normal vibrational modes.

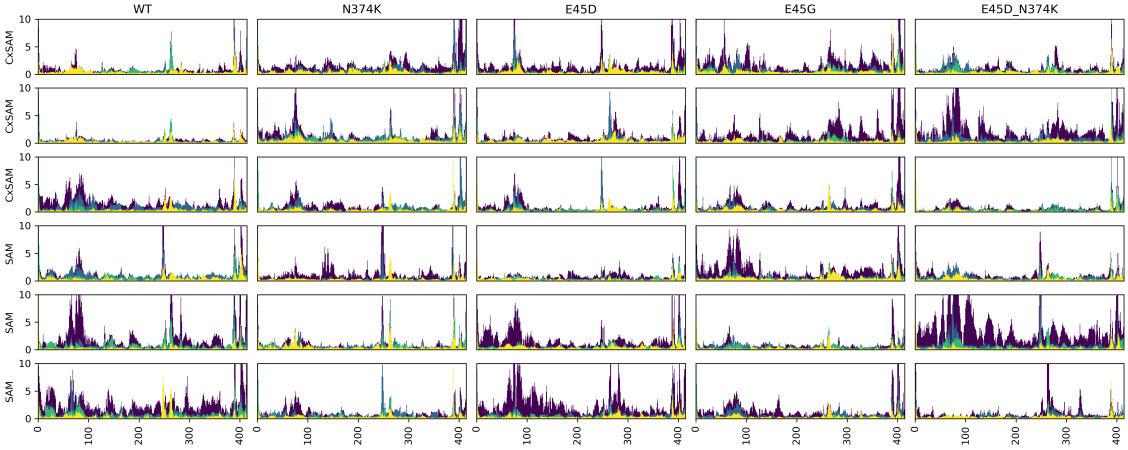


FIGURE F.4. First four (4) normal vibrational modes with contributions by each residue to each mode, shown for all variant/cosubstrate combinations in triplicate.

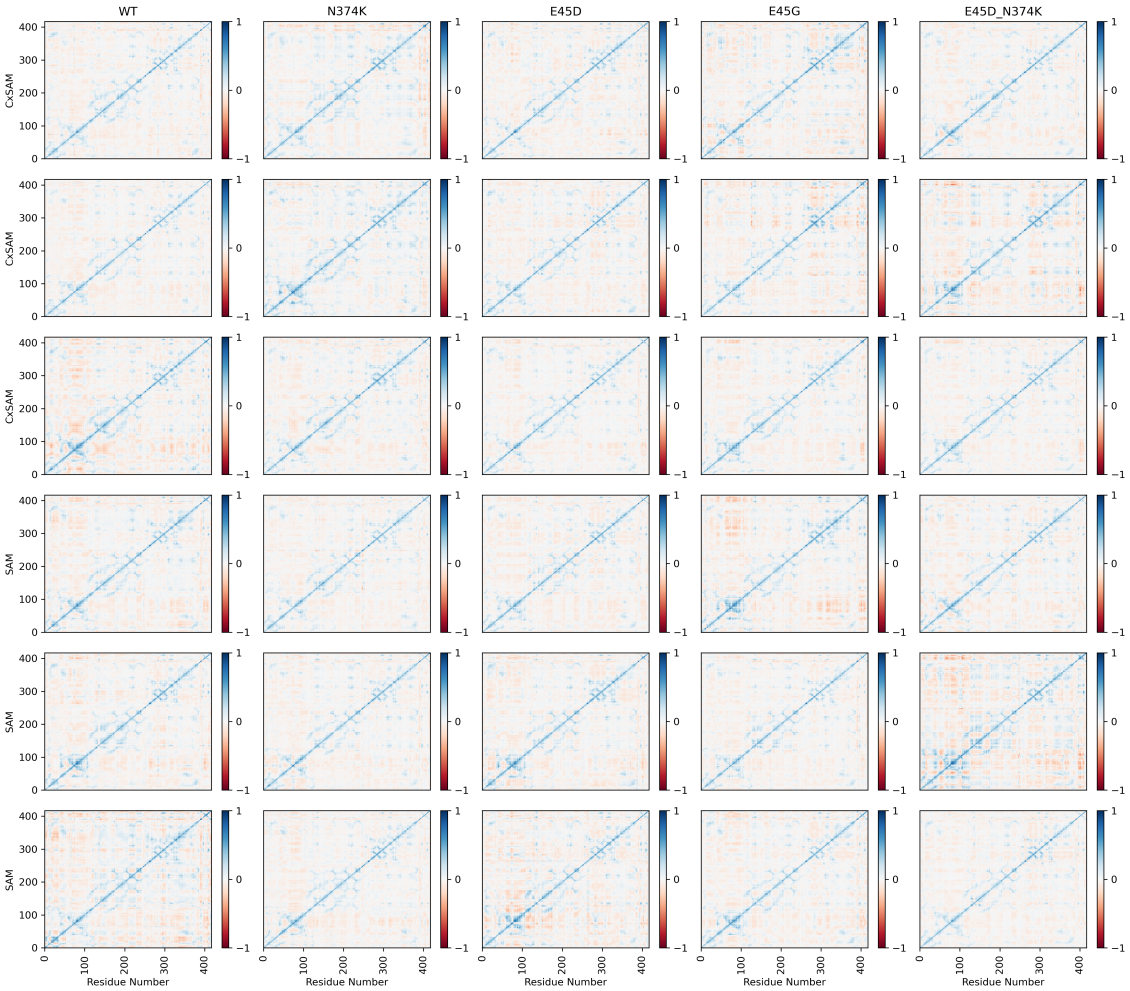


FIGURE F.5. Correlated motion for all variant/cosubstrate combinations in triplicate. Correlated motion shows the relationship of motion between pairs of residues for all pairs, with negative values in red showing strongly anticorrelated motion and positive values in blue showing strongly correlated motion.

REFERENCES

[1] Mark A. Hix and G. Andrés Cisneros, *Computational Investigation of APOBEC3H Substrate Orientation and Selectivity*, Journal of Physical Chemistry B 124 (2020), no. 19, 3903–3908.

[2] Mark A Hix, Lai Wong, Ben Flath, Linda Chelico, and G. Andrés Cisneros, *Single-nucleotide polymorphism of the DNA cytosine deaminase APOBEC3H haplotype I leads to enzyme destabilization and correlates with lung cancer*, NAR Cancer 2 (2020), no. 3, 1–27.

[3] Amit Gaba, Mark A. Hix, Sana Suhail, Ben Flath, Brock Boysan, Danielle R. Williams, Tomas Pelletier, Michael Emerman, Faruck Morcos, G. Andrés Cisneros, and Linda Chelico, *Divergence in dimerization and activity of primate apobec3c*, bioRxiv (2021).

[4] Jessica Lohrman, Erik A. Vázquez-Montelongo, Subhamay Pramanik, Victor W. Day, Mark A. Hix, Kristin Bowman-James, and G. Andrés Cisneros, *Characterizing Hydrogen-Bond Interactions in Pyrazinetetracarboxamide Complexes: Insights from Experimental and Quantum Topological Analyses*, Inorganic Chemistry 57 (2018), no. 16, 9775–9778.

[5] Mark A. Hix, Emmett M. Leddin, and G. Andrés Cisneros, *Combining evolutionary conservation and quantum topological analyses to determine quantum mechanics subsystems for biomolecular quantum mechanics/Molecular mechanics simulations*, Journal of Chemical Theory and Computation 17 (2021), no. 7, 4524–4537.

[6] D. Morales-Morales and C.M. Jensen (eds.), *The Chemistry of Pincer Compounds*, Elsevier, Amsterdam, The Netherlands, 2007.

[7] Stephen J. Loeb and George K. H. Shimizu, *Dimetalated thioether complexes as building blocks for organometallic coordination polymers and aggregates*, J. Chem. Soc., Chem. Commun. (1993), no. 18, 1395.

[8] Pablo Steenwinkel, Huub Kooijman, Wilberth J. J. Smeets, Anthony L. Spek, David M. Grove, and Gerard Van Koten, *Intramolecularly stabilized 1,4-phenylene-bridged homo- and heterodinuclear palladium and platinum organometallic complexes containing n,c,n-*

coordination motifs; η1-so2 coordination and formation of an organometallic arenium ion complex with two pt-c σ-bonds, Organometallics 17 (1998), no. 24, 5411–5426.

[9] Warren W. Gerhardt, Anthony J. Zucchero, James N. Wilson, Clinton R. South, Uwe H. F. Bunz, and Marcus Weck, *Supramolecular cruciforms*, Chem. Commun. (Cambridge, U. K.) (2006), no. 20, 2141–2143.

[10] Heng Zhang and Aiwen Lei, *Palladium(iv) chemistry supported by pincer type ligands*, Dalton Trans. 40 (2011), no. 35, 8745–8754.

[11] Jessica Lohrman, Hanumaiah Telikepalli, Thomas S. Johnson, Timothy A. Jackson, Victor W. Day, and Kristin Bowman-James, *Pyrazinetetracarboxamide: A duplex ligand for palladium(ii)*, Inorg. Chem. 55 (2016), no. 11, 5098–5100.

[12] Everly B. Fleischer and Mike B. Lawson, *Uni- and bimetallic complexes derived from a substituted pyrazine ligand*, Inorg. Chem. 11 (1972), no. 11, 2772.

[13] Everly B. Fleischer, David Jeter, and Roxanne Florian, *Structure of the dimetallic complex catena-μ-chloro-dichloro-μ-[n,n'-bis[2-(2-pyridyl)ethyl]-2,3-pyrazinedicarboxamidato-n,nn,n1:n',nn,n4-dicopper*, Inorg. Chem. 13 (1974), no. 5, 1042.

[14] Ipsita Mallik and Sanku Mallik, *Design and synthesis of new ligands for positioning two metal ions*, Synlett (1996), no. 8, 734–736.

[15] Julia Hausmann, Geoffrey B. Jameson, and Sally Brooker, *Control of molecular architecture by the degree of deprotonation: self-assembled di- and tetranuclear copper(ii) complexes of n,n'-bis(2-pyridylmethyl)pyrazine-2,3-dicarboxamide*, Chem. Commun. (Cambridge, U. K.) (2003), no. 24, 2992–2993.

[16] Dilovan S. Cati, Joan Ribas, Jordi Ribas-Arino, and Helen Stoeckli-Evans, *Self-assembly of cuii and niii [2 × 2] grid complexes and a binuclear cuii complex with a new semiflexible substituted pyrazine ligand: Multiple anion encapsulation and magnetic properties*, Inorg. Chem. 43 (2004), no. 3, 1021–1030.

[17] A. D. Becke and K. E. Edgecombe, *A simple measure of electron localization in atomic and molecular systems*, J. Chem. Phys. 92 (1990), no. 9, 5397.

[18] B. Silvi and A. Savin, *Classification of chemical bonds based on topological analysis of electron localization functions*, Nature (London) 371 (1994), no. 6499, 683.

[19] Bernard Silvi, *How topological partitions of the electron distributions reveal delocalization*, Phys. Chem. Chem. Phys. 6 (2004), no. 2, 256–260.

[20] Kenneth A Johnson, *The kinetic and chemical mechanism of high-fidelity dna polymerases.*, Biochim. Biophys. Acta 1804 (2010), no. 5, 1041–8.

[21] Julia Contreras-Garcia, Erin R. Johnson, Shahar Keinan, Robin Chaudret, Jean-Philip Piquemal, David N. Beratan, and Weitao Yang, *Nciplot: A program for plotting noncovalent interaction regions*, J. Chem. Theory Comput. 7 (2011), no. 3, 625–632.

[22] R.F.W Bader, *Atoms in Molecules: A Quantum Theory*, 1st ed., Oxford University Press, Oxford, U.K., 1994.

[23] Richard F. W. Bader, *A quantum theory of molecular structure and its applications*, Chem. Rev. 91 (1991), no. 5, 893–928.

[24] Richard F. W. Bader, *The quantum mechanical basis of conceptual chemistry*, Monatsh. Chem. 136 (2005), no. 6, 819–854.

[25] U. Koch and P. L. A. Popelier, *Characterization of c-h-o hydrogen bonds on the basis of the charge density*, J. Phys. Chem. 99 (1995), no. 24, 9747.

[26] P. L. A. Popelier, *Characterization of a dihydrogen bond on the basis of the electron density*, J. Phys. Chem. A 102 (1998), no. 10, 1873–1878.

[27] E. Espinosa, E. Molins, and C. Lecomte, *Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities*, Chem. Phys. Lett. 285 (1998), no. 3,4, 170–173.

[28] Robin Chaudret, G. Andres Cisneros, Olivier Parisel, and Jean-Philip Piquemal, *Unraveling low-barrier hydrogen bonds in complex systems with a simple quantum topological criterion*, Chem. - Eur. J. 17 (2011), no. 10, 2833.

[29] J.-P. Piquemal, J. Pilme, O. Parisel, H. Gerard, I. Fourre, J. Berges, C. Gourlaouen, A. De La Lande, M.-C. Van Severen, and B. Silvi, *What can be learnt on biologically relevant*

systems from the topological analysis of the electron localization function?, Int. J. Quantum Chem. 108 (2008), no. 11, 1951–1969.

[30] Robin Chaudret, Jean-Philip Piquemal, and G. Andres Cisneros, *Correlation between electron localization and metal ion mutagenicity in dna synthesis from qm/mm calculations*, Phys. Chem. Chem. Phys. 13 (2011), no. 23, 11239–11247.

[31] James A. Bellow, Dong Fang, Natalija Kovacevic, Philip D. Martin, Jason Shearer, G. Andres Cisneros, and Stanislav Groysman, *Novel alkoxide cluster topologies featuring rare seesaw geometry at transition metal centers*, Chem. - Eur. J. 19 (2013), no. 37, 12225–12228.

[32] Dong Fang, Robin Chaudret, Jean-Philip Piquemal, and G. Andres Cisneros, *Toward a deeper understanding of enzyme reactions using the coupled elf/nci analysis: Application to dna repair enzymes*, J. Chem. Theory Comput. 9 (2013), no. 5, 2156–2160.

[33] Nipuni-Dhanesha H. Gamage, Benedikt Stiasny, Eric G. Kratz, Jorg Stierstorfer, Philip D. Martin, Gerardo Andres Cisneros, Thomas M. Klapotke, and Charles Hartger Winter, *Energetic materials trends in 5- and 6-membered cyclic peroxides containing hydroperoxy and hydroxy substituents*, Eur. J. Inorg. Chem. 2016 (2016), no. 31, 5036–5043.

[34] Dennis Madsen, Claus Flensburg, and Sine Larsen, *Properties of the experimental crystal charge density of methylammonium hydrogen maleate. a salt with a very short intramolecular o-h-o hydrogen bond*, J. Phys. Chem. A 102 (1998), no. 12, 2177–2188.

[35] M. V. Vener, A. V. Manaev, A. N. Egorova, and V. G. Tsirelson, *Qtaim study of strong h-bonds with the o-h···a fragment (a = o, n) in three-dimensional periodical crystals*, J. Phys. Chem. A 111 (2007), no. 6, 1155–1162.

[36] Yirong Mo, *Can qtaim topological parameters be a measure of hydrogen bonding strength?*, J. Phys. Chem. A 116 (2012), no. 21, 5240–5246.

[37] William D. Arnold and Eric Oldfield, *The chemical nature of hydrogen bonding in proteins via nmr: J-couplings, chemical shifts, and aim theory*, J. Am. Chem. Soc. 122 (2000), no. 51, 12835–12841.

[38] Chirag D. Gheewala, Bridget E. Collins, and Tristan H. Lambert, *An aromatic ion*

platform for enantioselective bronsted acid catalysis, Science (Washington, DC, U. S.) 351 (2016), no. 6276, 961–965.

[39] Rommie E. Amaro and Adrian J. Mulholland, *Multiscale methods in drug design bridge chemical and biological complexity in the search for cures*, Nat. Rev. Chem. 2 (2018).

[40] Hans Martin Senn and Walter Thiel, *QM/MM methods for biomolecular systems*, Angewandte Chemie - International Edition 48 (2009), no. 7, 1198–1229.

[41] Gerrit Groenhof, *Introduction to QM/MM simulations*, Methods in Molecular Biology 924 (2013), 43–66.

[42] Valerie Vaissier Welborn and Teresa Head-Gordon, *Fluctuations of Electric Fields in the Active Site of the Enzyme Ketosteroid Isomerase*, J. Am. Chem. Soc. 141 (2019), no. 32, 12487–12492.

[43] Jacek Dziedzic, Teresa Head-Gordon, Martin Head-Gordon, and Chris Kriton Skylaris, *Mutually polarizable QM/MM model with in situ optimized localized basis functions*, J. Chem. Phys. 150 (2019), no. 7.

[44] Iván Solt, Petr Kulhánek, Istvń Simon, Steven Winfield, Mike C. Payne, Gábor Csányi, and Monika Fuxreiter, *Evaluating boundary dependent errors in qm/mm simulations*, J. Phys. Chem. B.

[45] Denis Flaig, Matthias Beer, and Christian Ochsenfeld, *Convergence of electronic structure with the size of the qm region: Example of qm/mm nmr shieldings*, J. Chem. Theo. Comp.

[46] Rong-Zhen Liao and Walter Thiel, *Convergence in the qm-only and qm/mm modeling of enzymatic reactions: A case study for acetylene hydratase*, J. Comput. Chem.

[47] J.D. Hartman, T.J. Neubauer, B.G. Caulkins, L.J Mueller, and Beran G.J.O., *Converging nuclear magnetic shielding calculations with respect to basis and system size in protein systems*, J. Biomol. NMR 62 (2015), no. 7, 327–340.

[48] Heather J. Kulik, Jianyu Zhang, Judith P. Klinman, and Todd J. Martínez, *How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O - Methyltransferase*, The Journal of Physical Chemistry B 120 (2016), no. 44, 11381–11394.

[49] Garima Jindal and Arieh Warshel, *Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region*, Journal of Physical Chemistry B 120 (2016), no. 37, 9913–9921.

[50] Susanta Das, Kwangho Nam, and Dan Thomas Major, *Rapid convergence of energy and free energy profiles with quantum mechanical size in quantum mechanical–molecular mechanical simulations of proton transfer in dna*, J. Chem. Theo. Comp.

[51] Qiang Cui, Tanmoy Pal, and Luke Xie, *Biomolecular qm/mm simulations: What are some of the burning issues?*, J. Phys. Chem. B 0 (2021), no. 0, null.

[52] Sophie Sumner, Pär Söderhjelm, and Ulf Ryde, *Effect of geometry optimizations on qm-cluster and qm/mm studies of reaction energies in proteins*, J. Chem. Theo Comput. 9 (2013), no. 9, 4205–4214.

[53] Jógvan Magnus Haugaard Olsen, Nanna Holmgaard List, Kasper Kristensen, and Jacob Kongsted, *Accuracy of protein embedding potentials: An analysis in terms of electrostatic potentials*, J. Chem. Theo. Comput. 11 (2015), no. 4, 1832–1842.

[54] Eric G. Kratz, Alice R. Walker, Louis Lagardère, Filippo Lipparini, Jean-Philip Piquemal, and G. Andrés Cisneros, *Lichem: A qm/mm program for simulations with multipolar and polarizable force fields*, J. Comput. Chem.

[55] Daniele Loco, Louis Lagardère, Stefano Caprasecca, Filippo Lipparini, Benedetta Mennucci, and Jean-Philip Piquemal, *Hybrid qm/mm molecular dynamics with amoeba polarizable embedding*, J. Chem. Theo. Comput. 13 (2017), no. 9, 4025–4033.

[56] Lina J. Nabo, Jógvan Magnus Haugaard Olsen, Todd J. Martínez, and Jacob Kongsted, *The quality of the embedding potential is decisive for minimal quantum region size in embedding calculations: The case of the green fluorescent protein*, J. Chem. Theo. Comput. 13 (2017), no. 12, 6230–6236.

[57] Daniele Loco, Louis Lagardère, G. Andrés. Cisneros, Giovanni Scalmani, Michael Frisch, Filippo Lipparini, Benedetta Mennucci, and Jean-Philip Piquemal, *Towards large scale hybrid qm/mm dynamics of complex systems with advanced point dipole polarizable embeddings*, Chem. Sci. 10 (2019), 7200–7211.

[58] Jorge Nochebuena, Sehr Naseem-Khan, and G. Andrés Cisneros, *Development and application of qm/mm methods with advanced polarizable potentials*, 2020.

[59] G. Andrés Cisneros, Min Wang, Peter Silinski, Michael C. Fitzgerald, and Weitao Yang, *Theoretical and experimental determination on two substrates turned over by 4-oxalocrotonate tautomerase*, Journal of Physical Chemistry A 110 (2006), no. 2, 700–708.

[60] G. Andrés Cisneros, Haiyan Liu, Yingkai Zhang, and Weitao Yang, *Ab initio QM/MM study shows there is no general acid in the reaction catalyzed by 4-oxalocrotonate tautomerase*, Journal of the American Chemical Society 125 (2003), no. 34, 10384–10393.

[61] Pan Wu, G. Andrés Cisneros, Hao Hu, Robin Chaudret, Xiangqian Hu, and Weitao Yang, *Catalytic mechanism of 4-oxalocrotonate tautomerase: Significances of protein - Protein interactions on proton transfer pathways*, Journal of Physical Chemistry B 116 (2012), no. 23, 6889–6897.

[62] Hedieh Torabifard and G. Andrés Cisneros, *Insight into wild-type and T1372E TET2-mediated 5hmC oxidation using ab initio QM/MM calculations*, Chemical Science 9 (2018), no. 44, 8433–8445.

[63] G. Andrés Cisneros, Lalith Perera, Miguel García-Díaz, Katarzyna Bebenek, Thomas A. Kunkel, and Lee G. Pedersen, *Catalytic mechanism of human DNA polymerase λ with Mg2+ and Mn2+ from ab initio quantum mechanical/molecular mechanical studies*, DNA Repair 7 (2008), no. 11, 1824–1834.

[64] Robin Chaudret, Jean Philip Piquemal, and G. Andrés Cisneros, *Correlation between electron localization and metal ion mutagenicity in DNA synthesis from QM/MM calculations*, Physical Chemistry Chemical Physics 13 (2011), no. 23, 11239–11247.

[65] Lorenzo Casalino, Łukasz Nierzwicki, Martin Jinek, and Giulia Palermo, *Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR-Cas9 Revealed by Ab Initio Molecular Dynamics*, ACS Catalysis (2020), 13596–13605.

[66] Andrew C. Pratt, Sajeewa W. Dewage, Allan H. Pang, Tapan Biswas, Sandra Barnard-Britson, G. Andrés Cisneros, and Oleg V. Tsodikov, *Structural and computational dissection*

of the catalytic mechanism of the inorganic pyrophosphatase from Mycobacterium tuberculosis, Journal of Structural Biology 192 (2015), no. 1, 76–87.

[67] Dong Fang and G. Andrés Cisneros, *Alternative pathway for the reaction catalyzed by dna dealkylase alkb from ab initio qm/mm calculations*, Journal of Chemical Theory and Computation 10 (2014), no. 11, 5136–5148.

[68] G. Andrés Cisneros, Lalith Perera, Roel M. Sehaaper, Lars C. Pedersen, Robert E. London, Lee G. Pedersen, and Thomas A. Darden, *Reaction mechanism of the ϵ subunit of E. coli DNA polymerase III: Insights into active site metal coordination and catalytically significant residues*, Journal of the American Chemical Society 131 (2009), no. 4, 1550–1556.

[69] Iván Solt, Petr Kulhánek, István Simon, Steven Winfield, Mike C. Payne, Gábor Csányi, and Monika Fuxreiter, *Evaluating boundary dependent errors in QM/MM simulations*, Journal of Physical Chemistry B 113 (2009), no. 17, 5728–5735.

[70] Nathan J. DeYonker and Charles Edwin Webster, *A Theoretical Study of Phosphoryl Transfers of Tyrosyl-DNA Phosphodiesterase i (Tdp1) and the Possibility of a "dead-End" Phosphohistidine Intermediate*, Biochemistry 54 (2015), no. 27, 4236–4247.

[71] Helena W. Qi, Maria Karelina, and Heather J. Kulik, *Quantifying electronic effects in QM and QM/MM biomolecular modeling with the Fukui function*, Wuli Huaxue Xuebao/ Acta Physico - Chimica Sinica 34 (2017), no. 1, 81–91.

[72] Maria Karelina and Heather J. Kulik, *Systematic Quantum Mechanical Region Determination in QM/MM Simulation*, Journal of Chemical Theory and Computation 13 (2017), no. 2, 563–576.

[73] Jack Fuller, Tim R. Wilson, Mark E. Eberhart, and Anastassia N. Alexandrova, *Charge Density in Enzyme Active Site as a Descriptor of Electrostatic Preorganization*, Journal of Chemical Information and Modeling 59 (2019), no. 5, 2367–2373.

[74] Aurélien De La Lande, Aurelio Alvarez-Ibarra, Karim Hasnaoui, Fabien Cailliez, Xiaojing Wu, Tzonka Mineva, Jérôme Cuny, Patrizia Calaminici, Luis López-Sosa, Gerald Geudtner, Isabelle Navizet, Cristina Garcia Iriepa, Dennis R. Salahub, and Andreas M.

Köster, *Molecular simulations with in-deMon2k QM/MM, a tutorial-review*, Molecules 24 (2019), no. 9.

[75] Monica Yun Liu, Hedieh Torabifard, Daniel J. Crawford, Jamie E. DeNizio, Xing Jun Cao, Benjamin A. Garcia, G. Andrés Cisneros, and Rahul M. Kohli, *Mutations along a TET2 active site scaffold stall oxidation at 5-hydroxymethylcytosine*, Nature Chemical Biology 13 (2017), no. 2, 181–187.

[76] A.J.M. Ribeiro, J.D. Tyzack, N. Borkakoti, G.M. Holliday, and Thornton J.M., *A global analysis of function and conservation of catalytic residues in enzymes*, J. Biol. Chem. 295 (2020), 314–324.

[77] Frances H. Arnold, *Directed evolution: Bringing new chemistry to life*, Ang. Chem. Intl. Ed. 57 (2018), no. 16, 4143–4148.

[78] In Geol Choi and Sung Hou Kim, *Evolution of protein structural classes and protein sequence families*, Proceedings of the National Academy of Sciences of the United States of America 103 (2006), no. 38, 14056–14061.

[79] Anna R. Panchenko and Thomas Madej, *Structural similarity of loops in protein families: Toward the understanding of protein evolution*, BMC Evolutionary Biology 5 (2005), 1–8.

[80] Barbara Spellerberg, Simone Martin, Josephine Weber-Heynemann, Norbert Schnitzler, Rudolf Lütticken, Eva Rozdzinski, and Andreas Podbielski, *Lmb, a protein with similarities to the LraI adhesin family, mediates attachment of Streptococcus agalactiae to human laminin*, Infection and Immunity 67 (1999), no. 2, 871–878.

[81] 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 (2016), no. 7, 578–594.

[82] Iddo Friedberg and Hanah Margalit, *Persistently conserved positions in structurally similar, sequence dissimilar proteins: Roles in preserving protein fold and function*, Protein Science 11 (2009), no. 2, 350–360.

[83] Christian P. Whitman, *The 4-oxalocrotonate tautomerase family of enzymes: How na-*

ture makes new enzymes using a β-α-β structural motif, Archives of Biochemistry and Biophysics 402 (2002), no. 1, 1–13.

[84] Naveenan Navaratnam and Rizwan Sarwar, *An overview of cytidine deaminases*, International Journal of Hematology 83 (2006), no. 3, 195–200.

[85] J.-P. Piquemal, J. Pilmé, O. Parisel, H. Gérard, I. Fourré, J. Bergès, C. Gourlaouen, A. De La Lande, M.-C. Van Severen, and B. Silvi, *What can be learnt on biologically relevant systems from the topological analysis of the electron localization function?*, Intl. J. Quant. Chem. 108, no. 11, 1951–1969.

[86] Axel D. Becke and K. E. Edgecombe, *A simple measure of electron localization in atomic and molecular systems.*

[87] Bernard Silvi and Andreas Savin, *Classification of chemical bonds based on topological analysis of electron localization functions*, Nature.

[88] Natacha Gillet, Robin Chaudret, Julia Contreras-García, Weitao Yang, Bernard Silvi, and Jean-Philip Piquemal, *Coupling quantum interpretative techniques: Another look at chemical mechanisms in organic reactions*, J. Chem. Theo. Comp.

[89] Dong Fang, Robin Chaudret, Jean-Philip Piquemal, and G. Andrés Cisneros, *Toward a deeper understanding of enzyme reactions using the coupled elf/nci analysis: Application to dna repair enzymes*, J. Chem. Theo. Comp.

[90] Julien Pilme and Jean-Philip Piquemal, *Advancing beyond charge analysis using the electronic localization function: Chemically intuitive distribution of electrostatic moments*, J. Comput. Chem.

[91] C. Notredame, D. G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment*, J Mol Biol 302 (2000), no. 1, 205–217.

[92] O. O'Sullivan, K. Suhre, C. Abergel, D. G. Higgins, and C. Notredame, *3DCoffee: combining protein sequences and structures within multiple sequence alignments*, J Mol Biol 340 (2004), no. 2, 385–395.

[93] O. Poirot, K. Suhre, C. Abergel, E. O'Toole, and C. Notredame, *3DCoffee@igs: a web*

server for combining sequences and structures into a multiple sequence alignment, Nucleic Acids Res 32 (2004), no. Web Server issue, 37–40.

[94] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame, *Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee*, Nucleic Acids Res 34 (2006), no. Web Server issue, W604–608.

[95] P. Di Tommaso, S. Moretti, I. Xenarios, M. Orobitg, A. Montanyola, J. M. Chang, J. F. Taly, and C. Notredame, *T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension*, Nucleic Acids Res 39 (2011), no. Web Server issue, W13–17.

[96] S. Moretti, F. Armougom, I. M. Wallace, D. G. Higgins, C. V. Jongeneel, and C. Notredame, *The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods*, Nucleic Acids Res 35 (2007), no. Web Server issue, W645–648.

[97] I. M. Wallace, O. O'Sullivan, D. G. Higgins, and C. Notredame, *M-Coffee: combining multiple sequence alignment methods with T-Coffee*, Nucleic Acids Res 34 (2006), no. 6, 1692–1699.

[98] Alex Bateman, *UniProt: A worldwide hub of protein knowledge*, Nucleic Acids Res. 47 (2019), no. D1, D506–D515.

[99] G. Andrés Cisneros, Min Wang, Peter Silinski, Michael C. Fitzgerald, and Weitao Yang, *The protein backbone makes important contributions to 4-oxalocrotonate tautomerase enzyme catalysis: Understanding from theory and experiment*, Biochemistry 43 (2004), no. 22, 6885–6892.

[100] Alice R. Walker and G. Andrés Cisneros, *Computational Simulations of DNA Polymerases: Detailed Insights on Structure/Function/Mechanism from Native Proteins to Cancer Variants*, Chemical Research in Toxicology 30 (2017), no. 11, 1922–1935.

[101] Alexander B. Taylor, Robert M. Czerwinski, William H. Johnson Jr., Christian P. Whitman, and Marvin L. Hackert, *Crystal structure of 4–oxalocrotonate tautomerase inac-*

tivated by 2–oxo–3–pentynoate at 2.4 åresolution: Analysis and implications for the mechanism of inactivation and catalysis, Biochemistry.

[102] Lulu Hu, Ze Li, Jingdong Cheng, Qinhui Rao, Wei Gong, Mengjie Liu, Yujiang Geno Shi, Jiayu Zhu, Ping Wang, and Yanhui Xu, *Crystal structure of tet2-dna complex: Insight into tet-mediated 5mc oxidation*, Cell 155 (2013), no. 7, 1545–1555.

[103] Miguel García-Díaz, Katarzyna Bebenek, Rosario Sabariegos, Orlando Domínguez, Josana Rodríguez, Tomas Kirchhoff, Esther García-Palomero, Angel J. Picher, Raquel Juárez, Jose F. Ruiz, Thomas A. Kunkel, and Luis Blanco, *DNA Polymerase λ, a Novel DNA Repair Enzyme in Human Cells*, J. Biol. Chem.

[104] Miguel Garcia-Diaz, Katarzyna Bebenek, Joseph M Krahn, Luis Blanco, Thomas A Kunkel, and Lars C Pedersen, *A structural solution for the dna polymerase λ-dependent repair of dna gaps with minimal homology*, Molecular Cell 13 (2004), no. 4, 561–572.

[105] Chuan Tian, Koushik Kasavajhala, Kellon A. A. Belfon, Lauren Raguette, He Huang, Angela N. Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling, *ff19sb: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution*, J. .Chem. Theo. .Comput. 16 (2020), no. 1, 528–552.

[106] Eric G Kratz, Alice R Walker, Louis Lagardère, Filippo Lipparini, Jean Philip Piquemal, and G. Andrés Cisneros, *LICHEM: A QM/MM program for simulations with multipolar and polarizable force fields*, J. Comput. Chem. 37 (2016), no. 11, 1019–1029.

[107] Hatice Gokcan, Erik A. Vázquez-Mongelongo, and G. Andrés Cisneros, *Lichem 1.1: Recent improvements and new capabilities*, J. Chem. Theo. Comp.

[108] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega,

J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, , and D. J. Fox, *Gaussian 16, revision a.03*, Gaussian, Inc., Wallingford, CT, 2017.

[109] J.W. Ponder, *Tinker, software tools for molecular design, version 5.0: the most updated version for the tinker program can be obtained from j.w. ponder's www site at http://dasher.wustl.edu/tinker.*, Washington University, St. Louis, 2004.

[110] Stéphane Noury, Xénophon Krokidis, Franck Fuster, and Bernard Silvi, *Computational tools for the electron localization function topological analysis*, Comput. Chem. 23 (1999), no. 6, 597–604.

[111] Tian Lu and Feiwu Chen, *Multiwfn: A multifunctional wavefunction analyzer*, Journal of Computational Chemistry 33 (2012), no. 5, 580–592.

[112] William Humphrey, Andrew Dalke, and Klaus Schulten, *VMD: Visual molecular dynamics*, Journal of Molecular Graphics 14 (1996), no. 1, 33 – 38.

[113] 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 (2004), no. 13, 1605–1612.

[114] Tyler M. M. Stack, Wenzong Li, William H. Johnson, Yan Jessie Zhang, and Christian P. Whitman, *Inactivation of 4-oxalocrotonate tautomerase by 5-halo-2-hydroxy-2,4-pentadienoates*, Biochem. 57 (2018), no. 6, 1012–1021.

[115] H. F. Azurmendi, S. G. Miller, C. P. Whitman, and A. S. Mildvan, *Half–of–the–sites binding of reactive intermediates and their analogues to 4–oxalocrotonate tautomerase and induced structural assymetry of the enzyme*, Biochemistry.

[116] Thomas K. Harris, Robert M. Czerwinski, William H. Johnson, Patricia M. Legler, Chitrananda Abeygunawardana, Michael A. Massiah, James T. Stivers, Christian P. Whit-

man, and Albert S. Mildvan, *Kinetic, stereochemical, and structural effects of mutations of the active site arginine residues in 4-oxalocrotonate tautomerase*, Biochemistry 38 (1999), no. 38, 12343–12357.

[117] Tell Tuttle, Ehud Keinan, and Walter Thiel, *Understanding the enzymatic activity of 4-oxalocrotonate tautomerase and its mutant analogues: a computational study.*, J. Phys. Chem. B.

[118] Tell Tuttle and Walter Thiel, *Substrate orientation in 4–oxalocrotonate tautomerase and its effect on qm/mm energy profiles*, J. Phys. Chem. B.

[119] Elizabeth A. Burks, Christopher D. Fleming, Andrew D. Mesecar, Christian P. Whitman, and Scott D. Pegan, *Kinetic and structural characterization of a heterohexamer 4-oxalocrotonate tautomerase from chloroflexus aurantiacus J-10-fl: Implications for functional and structural diversity in the tautomerase superfamily*, Biochemistry 49 (2010), no. 24, 5016–5027.

[120] N. Metanis, A. Brik, P. E. Dawson, and E. Keinan, *Electrostatic interactions dominate the catalytic contribution of arg39 in 4–oxalocrotonate tautomerase*, J. Am. Chem. Soc.

[121] James D. R. Knight, Donald Hamelberg, J. Andrew McCammon, and Rashmi Kothary, *The role of conserved water molecules in the catalytic domain of protein kinases*, Proteins: Structure, Function, and Bioinformatics 76 (2009), no. 3, 527–535.

[122] Elena Decaneto, Tatiana Vasilevskaya, Yuri Kutin, Hideaki Ogata, Moran Grossman, Irit Sagi, Martina Havenith, Wolfgang Lubitz, Walter Thiel, and Nicholas Cox, *Solvent water interactions within the active site of the membrane type i matrix metalloproteinase*, Phys. Chem. Chem. Phys. 19 (2017), 30316–30331.

[123] Dong Fang, Richard L. Lord, and G. Andrés Cisneros, *Ab initio qm/mm calculations show an intersystem crossing in the hydrogen abstraction step in dealkylation catalyzed by alkb*, J. Phys. Chem. B.

[124] Mary E. Anderson, Benoît Braïda, Philippe C. Hiberty, and Thomas R. Cundari, *Revealing a decisive role for secondary coordination sphere nucleophiles on methane activation*, J. Am. Chem. Soc. 142 (2020), no. 6, 3125–3131.

[125] Jenny G. Vitillo, Connie C. Lu, Christopher J. Cramer, Aditya Bhan, and Laura Gagliardi, *Influence of first and second coordination environment on structural fe(ii) sites in mil-101 for c–h bond activation in methane*, ACS Cat. 11 (2021), no. 2, 579–589.

[126] Jennifer Yamtich and Joann B Sweasy, *DNA polymerase Family X: Function, structure, and cellular roles.*, Biochimica et Biophysica Acta 1804 (2010), no. 5, 1136–50.

[127] Isabelle Frouin, Magali Toueille, Elena Ferrari, Igor Shevelev, and Ulrich Hübscher, *Phosphorylation of human DNA polymerase λ by the cyclin-dependent kinase Cdk2/cyclin A complex is modulated by its association with proliferating cell nuclear antigen*, Nucleic Acids Research 33 (2005), no. 16, 5354–5361.

[128] T A Steitz, *DNA polymerases: structural diversity and common mechanisms.*, The Journal of biological chemistry 274 (1999), no. 25, 17395–8.

[129] J. A. Cowan, *Structural and catalytic chemistry of magnesium-dependent enzymes*, Biometals 15 (2002), no. 3, 225–235.

[130] M. Garcia-Diaz, K. Bebenek, J. M. Krahn, L. Blanco, T. A. Kunkel, and L. C. Pedersen, *A structural solution for the DNA polymerase lambda-dependent repair of DNA gaps with minimal homology*, Mol Cell 13 (2004), no. 4, 561–572.

[131] Miguel Garcia-Diaz, Katarzyna Bebenek, Guanghua Gao, Lars C. Pedersen, Robert E. London, and Thomas A. Kunkel, *Structure–function studies of dna polymerase lambda*, DNA Repair 4 (2005), no. 12, 1358 – 1367.

[132] T. Nakamura, Y. Zhao, Y. Yamagata, Y. J. Hua, and W. Yang, *Watching DNA polymerase η make a phosphodiester bond*, Nature 487 (2012), no. 7406, 196–201.

[133] B. D. Freudenthal, W. A. Beard, D. D. Shock, and Wilson S. H., *Observing a DNA polymerase choose right from wrong.*, Cell 154 (2013), no. 1, 157–168.

[134] Yang Gao and Wei Yang, *Capture of a third mg2+ is essential for catalyzing dna synthesis*, Science 352 (2016), no. 6291, 1334–1337.

[135] Lalith Perera, Ulrich Essmann, and Max L. Berkowitz, *Effect of the treatment of long-range forces on the dynamics of ions in aqueous solutions.*

[136] Lalith Perera, Bret D. Freudenthal, William A. Beard, David D. Shock, Lee G. Ped-

ersen, and Samuel H. Wilson, *Requirement for transient metal ions revealed through computational analysis for dna polymerase going in reverse*, Proc. Natl. Acad. Sci.

[137] Lalith Perera, Bret D. Greudenthal, William A. Beard, Lee G. Pedersen, and Samuel H Wilson, *Revealing the role of the product metal in dna polymerase β catalysis*, Nucl. Ac. Res.

[138] David R. Stevens and Sharon Hammes-Schiffer, *Exploring the role of the third active site metal ion in dna polymerase η with qm/mm free energy simulations*, J. Am. Chem. Soc. 140 (2018), no. 28, 8965–8969.

[139] G. Andrés Cisneros, M. Wang, P. Silinski, M.C. Fitzgerald, and W. Yang, *The protein backbone makes important contributions to 4-oxalocrotonate tautomerase enzyme catalysis: Understanding from theory and experiment*, Biochemistry.

[140] Robin Chaudret, Jean-Philip Piquemal, and G. Andrés Cisneros, *Correlation between electron localization and metal ion mutagenicity in dna synthesis from qm/mm calculations*, Phys. Chem. Chem. Phys.

[141] Katarzyna Bebenek, Miguel Garcia-Diaz, Rui-Zhe Zhou, Lawrence F. Povirk, and Thomas A. Kunkel, *Loop 1 modulates the fidelity of DNA polymerase λ*, Nuc. Ac. Res. 38 (2010), no. 16, 5419–5431.

[142] Manjuan Liu, Aurélie Mallinger, Marcello Tortorici, Yvette Newbatt, Meirion Richards, Amin Mirza, Rob L.M. Van Montfort, Rosemary Burke, Julian Blagg, and Teresa Kaserer, *Evaluation of APOBEC3B recognition motifs by NMR reveals preferred substrates*, ACS Chemical Biology 13 (2018), no. 9, 2427–2432.

[143] 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, Rommie E Amaro, Daniel A Harki, Reuben S Harris, and Hideki Aihara, *Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B.*, Nature structural & molecular biology 24 (2017), no. 2, 131–139.

[144] Diako Ebrahimi, Hamid Alinejad-Rokny, and Miles P. Davenport, *Insights into the motif preference of APOBEC3 enzymes*, PLoS ONE 9 (2014), no. 1, 1–9.

[145] Silvestro G. Conticello, *The aid/apobec family of nucleic acid mutators*, Genome Biology 9 (2008), no. 6, 229.

[146] Gabriel J. Starrett, Elizabeth M. Luengas, Jennifer L. McCann, Diako Ebrahimi, Nuri A. Temiz, Robin P. Love, Yuqing Feng, Madison B. Adolph, Linda Chelico, Emily K. Law, Michael A. Carpenter, and Reuben S. Harris, *The DNA cytosine deaminase APOBEC3H haplotype i likely contributes to breast and lung cancer mutagenesis*, Nature Communications 7 (2016), 12918.

[147] Nadine M. Shaban, Ke Shi, Kate V. Lauer, Michael A. Carpenter, Christopher M. Richards, Daniel Salamango, Jiayi Wang, Michael W. Lopresti, Surajit Banerjee, Rena Levin-Klein, William L. Brown, Hideki Aihara, and Reuben S. Harris, *The Antiviral and Cancer Genomic DNA Deaminase APOBEC3H Is Regulated by an RNA-Mediated Dimerization Mechanism*, Molecular Cell 69 (2018), no. 1, 75–86.e9.

[148] Fumiaki Ito, Hanjing Yang, Xiao Xiao, Shu Xing Li, Aaron Wolfe, Brett Zirkle, Vagan Arutiunian, and Xiaojiang S. Chen, *Understanding the Structure, Multimerization, Subcellular Localization and mC Selectivity of a Genomic Mutator and Anti-HIV Factor APOBEC3H*, Scientific Reports 8 (2018), no. 1, 3763.

[149] Alex Bateman, MariaJesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Andrew Cowley, Alan Da Silva, Maurizio De Giorgi, Tunca Dogan, Francesco Fazzini, Leyla Garcia Castro, Luis Figueira, Penelope Garmiri, George Georghiou, Daniel Gonzalez, Emma Hatton-Ellis, Weizhong Li, Wudong Liu, Rodrigo Lopez, Jie Luo, Yvonne Lussi, Alistair MacDougall, Andrew Nightingale, Barbara Palka, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Alexandre Renaux, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Elena Speretta, Edward Turner, Nidhi Tyagi, Vladimir Volynkin, Tony Wardell, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie-Claude Blatter, Brigitte

Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard de Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Cathy H Wu, Cecilia N Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh, and Jian Zhang, *UniProt: the universal protein knowledgebase*, Nucleic Acids Research 45 (2017), no. D1, D158–D169.

[150] C. E. A. F. Schafmeister, W. S. Ross, and V. Romanovski, *LEAP*, 1995.

[151] D.A. Case, D.S. Cerutti, T.E. Cheatham III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, and G.H. Monard, *Amber16*, 2017.

[152] 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 (2015), no. 8, 3696–3713.

[153] Marie Zgarbová, Jiří Šponer, Michal Otyepka, Thomas E. Cheatham, Rodrigo Galindo-Murillo, and Petr Jurečka, *Refinement of the sugar-phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA*, Journal of Chemical Theory and Computation 11 (2015), no. 12, 5723–5736.

[154] Ilyas Yildirim, Harry A. Stern, Scott D. Kennedy, Jason D. Tubbs, and Douglas H.

Turner, *Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine*, Journal of Chemical Theory and Computation 6 (2010), no. 5, 1520–1531.

[155] 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 (1983), no. 2, 926–935.

[156] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, *Molecular dynamics with coupling to an external bath*, The Journal of Chemical Physics 81 (1984), no. 8, 3684–3690.

[157] Andreas W. Götz, Mark J. Williamson, Dong Xu, Duncan Poole, Scott Le Grand, and Ross C. Walker, *Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born*, Journal of Chemical Theory and Computation 8 (2012), no. 5, 1542–1555.

[158] Romelia Salomon-Ferrer, Andreas W. Götz, 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 (2013), no. 9, 3878–3888.

[159] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen, *A smooth particle mesh Ewald method*, The Journal of Chemical Physics 103 (1995), no. 19, 8577–8593.

[160] Andrej Šali and Tom L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints*, Journal of Molecular Biology 234 (1993), no. 3, 779 – 815.

[161] András Fiser, Richard Kinh Gian Do, and Andrej Šali, *Modeling of loops in protein structures*, Protein Science 9 (2000), no. 9, 1753–1773.

[162] Daniel R. Roe and Thomas E. Cheatham, *PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data*, Journal of Chemical Theory and Computation 9 (2013), no. 7, 3084–3095.

[163] G. Andrés Cisneros, M. Wang, P. Silinski, M.C. Fitzgerald, and W. Yang, *Theoretical*

and experimental determination on two substrates turned over by 4–oxalocrotonate tautomerase, J. Phys. Chem. A.

[164] Sajeewa Walimuni Dewage and G. Andrés Cisneros, *Computational analysis of ammonia transfer along two intramolecular tunnels in Staphylococcus aureus glutamine-dependent amidotransferase (GatCAB)*, The Journal of Physical Chemistry B 119 (2015), no. 9, 3669–3677.

[165] Angela A. Elias and G. Andrés Cisneros, *Chapter two - computational study of putative residues involved in DNA synthesis fidelity checking in Thermus aquaticus DNA polymerase I*, Biomolecular Modelling and Simulations (Tatyana Karabencheva-Christova, ed.), Advances in protein chemistry and structural biology, vol. 96, Academic Press, 2014, pp. 39 – 75.

[166] Sarah E. Graham, FatimaSultana Syeda, and G. Andrés Cisneros, *Computational prediction of residues involved in fidelity checking for DNA synthesis in DNA polymerase I*, Biochemistry 51 (2012), no. 12, 2569–2578.

[167] Eric C. Logue, Nicolin Bloch, Erica Dhuey, Ruonan Zhang, Ping Cao, Cecile Herate, Lise Chauveau, Stevan R. Hubbard, and Nathaniel R. Landau, *A dna sequence recognition loop on apobec3a controls substrate specificity*, PLOS ONE 9 (2014), no. 5, 1–10.

[168] Ke Shi, Michael A Carpenter, Kayo Kurahashi, Reuben S Harris, and Hideki Aihara, *Crystal structure of the DNA deaminase APOBEC3B catalytic domain.*, The Journal of biological chemistry 290 (2015), no. 47, 28120–30.

[169] Xiuxiu Lu, Tianlong Zhang, Zeng Xu, Shanshan Liu, Bin Zhao, Wenxian Lan, Chunxi Wang, Jianping Ding, and Chunyang Cao, *Crystal structure of DNA cytidine deaminase ABOBEC3G catalytic deamination domain suggests a binding mode of full-length enzyme to single-stranded DNA.*, The Journal of biological chemistry 290 (2015), no. 7, 4010–21.

[170] Shurong Hou, Tania V Silvas, Florian Leidner, Ellen A Nalivaika, Hiroshi Matsuo, Nese Kurt Yilmaz, and Celia A Schiffer, *Structural analysis of the active site and DNA binding of human cytidine deaminase APOBEC3B*, Journal of Chemical Theory and Computation 15 (2019), no. 1, 637–647.

[171] Anurag Rathore, Michael A. Carpenter, Özlem Demir, Terumasa Ikeda, Ming Li, Nadine M. Shaban, Emily K. Law, Dmitry Anokhin, William L. Brown, Rommie E. Amaro, and Reuben S. Harris, *The local dinucleotide preference of APOBEC3G can be altered from 5'-CC to 5'-TC by a single amino acid substitution*, Journal of Molecular Biology 425 (2013), no. 22, 4442–4454.

[172] Rahul M. Kohli, Robert W. Maul, Amy F. Guminski, Rhonda L. McClure, Kiran S. Gajula, Huseyin Saribasak, Moira A. McMahon, Robert F. Siliciano, Patricia J. Gearhart, and James T. Stivers, *Local sequence targeting in the AID/APOBEC family differentially impacts retroviral restriction and antibody diversification*, Journal of Biological Chemistry 285 (2010), no. 52, 40956–40964.

[173] Michael A. Carpenter, Erandi Rajagurubandara, Priyanga Wijesinghe, and Ashok S. Bhagwat, *Determinants of sequence-specificity within human AID and APOBEC3G*, DNA Repair 9 (2010), no. 5, 579–587.

[174] Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E. Felberg, David H. Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W. Gohara, Todd Dolinsky, Robert Konecny, David R. Koes, Jens Erik Nielsen, Teresa Head-Gordon, Weihua Geng, Robert Krasny, Guo-Wei Wei, Michael J. Holst, J. Andrew McCammon, and Nathan A. Baker, *Improvements to the apbs biomolecular solvation software suite*, Protein Science 27 (2018), no. 1, 112–128.

[175] Madison B Adolph, Robin P Love, Yuqing Feng, and Linda Chelico, *Erratum: Enzyme cycling contributes to efficient induction of genome mutagenesis by the cytidine deaminase APOBEC3B (Nucleic Acids Research (2017) 45:20 (11925–11940) DOI: 10.1093/nar/gkx832)*, 2018, pp. 12190–12191.

[176] H C Smith, *RNA binding to APOBEC deaminases; Not simply a substrate for C to U editing*, RNA Biol 14 (2017), no. 9, 1153–1165.

[177] S. Venkatesan, R. Rosenthal, N. Kanu, N. McGranahan, J. Bartek, S. A. Quezada, J. Hare, R. S. Harris, and C. Swanton, *Perspective: APOBEC mutagenesis in drug re-*

sistance and immune escape in HIV and cancer evolution, Annals of Oncology 29 (2018), no. 3, 563–572.

[178] L B Alexandrov, S Nik-Zainal, D C Wedge, S A Aparicio, S Behjati, A V Biankin, G R Bignell, N Bolli, A Borg, A L Borresen-Dale, S Boyault, B Burkhardt, A P Butler, C Caldas, H R Davies, C Desmedt, R Eils, J E Eyfjord, J A Foekens, M Greaves, F Hosoda, B Hutter, T Ilicic, S Imbeaud, M Imielinski, N Jager, D T Jones, D T Jones, S Knappskog, M Kool, S R Lakhani, C Lopez-Otin, S Martin, N C Munshi, H Nakamura, P A Northcott, M Pajic, E Papaemmanuil, A Paradiso, J V Pearson, X S Puente, K Raine, M Ramakrishna, A L Richardson, J Richter, P Rosenstiel, M Schlesner, T N Schumacher, P N Span, J W Teague, Y Totoki, A N Tutt, R Valdes-Mas, M M van Buuren, L van 't Veer, A Vincent-Salomon, N Waddell, L R Yates, Initiative Australian Pancreatic Cancer Genome, Icgc Breast Cancer Mmml-Seq Consortium, Icgc Breast Cancer Mmml-Seq Consortium, Icgc PedBrain, J Zucman-Rossi, P A Futreal, U McDermott, P Lichter, M Meyerson, S M Grimmond, R Siebert, E Campo, T Shibata, S M Pfister, P J Campbell, and M R Stratton, *Signatures of mutational processes in human cancer*, Nature 500 (2013), no. 7463, 415–421.

[179] R S Harris, J F Hultquist, and D T Evans, *The restriction factors of human immunodeficiency virus*, J Biol Chem 287 (2012), no. 49, 40875–40883 (eng).

[180] R Suspene, D Guetard, M Henry, P Sommer, S Wain-Hobson, and J P Vartanian, *Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo*, Proc Natl Acad Sci U S A 102 (2005), no. 23, 8321–8326 (eng).

[181] A Z Cheng, J Yockteng-Melgar, M C Jarvis, N Malik-Soni, I Borozan, M A Carpenter, J L McCann, D Ebrahimi, N M Shaban, E Marcon, J Greenblatt, W L Brown, L Frappier, and R S Harris, *Epstein-Barr virus BORF2 inhibits cellular APOBEC3B to preserve viral genome integrity*, Nat Microbiol 4 (2019), no. 1, 78–88 (eng).

[182] D J Salamango, J T Becker, J L McCann, A Z Cheng, O Demir, R E Amaro, W L Brown, N M Shaban, and R S Harris, *APOBEC3H Subcellular Localization Determinants Define Zipcode for Targeting HIV-1 for Restriction*, Mol Cell Biol 38 (2018), no. 23.

[183] R P Bennett, V Presnyak, J E Wedekind, and H C Smith, *Nuclear Exclusion of the*

HIV-1 host defense factor APOBEC3G requires a novel cytoplasmic retention signal and is not dependent on RNA binding, J Biol Chem 283 (2008), no. 12, 7320–7327 (eng).

[184] V B Soros, W Yonemoto, and W C Greene, *Newly synthesized APOBEC3G is incorporated into HIV virions, inhibited by HIV RNA, and subsequently activated by RNase H*, PLoS Pathog 3 (2007), no. 2, e15 (eng).

[185] T Matsuoka, T Nagae, H Ode, H Awazu, T Kurosawa, A Hamano, K Matsuoka, A Hachiya, M Imahashi, Y Yokomaku, N Watanabe, and Y Iwatani, *Structural basis of chimpanzee APOBEC3H dimerization stabilized by double-stranded RNA*, Nucleic Acids Res 46 (2018), no. 19, 10368–10379.

[186] Meng Zhu, Yuzhuo Wang, Cheng Wang, Wei Shen, Jia Liu, Liguo Geng, Yang Cheng, Juncheng Dai, Guangfu Jin, Hongxia Ma, Zhibin Hu, and Hongbing Shen, *The eQTL-missense polymorphisms of APOBEC3H are associated with lung cancer risk in a Han Chinese population*, Scientific Reports 5 (2015), 14969.

[187] Michael B Burns, Nuri A Temiz, and Reuben S Harris, *Evidence for APOBEC3B mutagenesis in multiple human cancers*, Nature Genetics 45 (2013), no. 9, 977–983.

[188] Michael B Burns, Lela Lackey, Michael A Carpenter, Anurag Rathore, Allison M Land, Brandon Leonard, Eric W Refsland, Delshanee Kotandeniya, Natalia Tretyakova, Jason B Nikas, Douglas Yee, Nuri A Temiz, Duncan E Donohue, Rebecca M McDougle, William L Brown, Emily K Law, and Reuben S Harris, *APOBEC3B is an enzymatic source of mutation in breast cancer.*, Nature 494 (2013), no. 7437, 366–70.

[189] K Chan, S A Roberts, L J Klimczak, J F Sterling, N Saini, E P Malc, J Kim, D J Kwiatkowski, D C Fargo, P A Mieczkowski, G Getz, and D A Gordenin, *An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers*, Nat Genet 47 (2015), no. 9, 1067–1072.

[190] James I. Hoopes, Luis M. Cortez, Tony M. Mertz, Ewa P. Malc, Piotr A. Mieczkowski, and Steven A. Roberts, *APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication*, Cell Reports 14 (2016), no. 6, 1273–1282.

[191] Marat D. Kazanov, Steven A. Roberts, Paz Polak, John Stamatoyannopoulos, Leszek J.

Klimczak, Dmitry A. Gordenin, and Shamil R. Sunyaev, *APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions*, Cell Reports 13 (2015), no. 6, 1103–1109.

[192] N Kanu, M A Cerone, G Goh, L P Zalmas, J Bartkova, M Dietzen, N McGranahan, R Rogers, E K Law, I Gromova, M Kschischo, M I Walton, O W Rossanese, J Bartek, R S Harris, S Venkatesan, and C Swanton, *DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer*, Genome Biol 17 (2016), no. 1, 185.

[193] Nicholas J. Haradhvala, Paz Polak, Petar Stojanov, Kyle R. Covington, Eve Shinbrot, Julian M. Hess, Esther Rheinbay, Jaegil Kim, Yosef E. Maruvka, Lior Z. Braunstein, Atanas Kamburov, Philip C. Hanawalt, David A. Wheeler, Amnon Koren, Michael S. Lawrence, and Gad Getz, *Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair*, Cell 164 (2016), no. 3, 538–549.

[194] V B Seplyarskiy, R A Soldatov, K Y Popadin, S E Antonarakis, G A Bazykin, and S I Nikolaev, *APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication*, Genome Res 26 (2016), no. 2, 174–182.

[195] M C Jarvis, D Ebrahimi, N A Temiz, and R S Harris, *Mutation Signatures Including APOBEC in Cancer Cell Lines*, JNCI Cancer Spectr 2 (2018), no. 1.

[196] M B Adolph, A Ara, and L Chelico, *APOBEC3 Host Restriction Factors of HIV-1 Can Change the Template Switching Frequency of Reverse Transcriptase*, J Mol Biol 431 (2019), no. 7, 1339–1352.

[197] Steven A. Roberts and Dmitry A. Gordenin, *Clustered and genome-wide transient mutagenesis in human cancers: Hypermutation without permanent mutators or loss of fitness*, BioEssays 36 (2014), no. 4, 382–393.

[198] Michael B. Burns, Brandon Leonard, and Reuben S. Harris, *APOBEC3B: Pathological consequences of an innate immune DNA mutator*, Biomedical Journal 38 (2015), no. 2, 102–110.

[199] Steven A. Roberts, Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargo, Petar Stojanov, Adam Kiezun, Gregory V. Kryukov, Scott L. Carter, Gordon Sak-

sena, Shawn Harris, Ruchir R. Shah, Michael A. Resnick, Gad Getz, and Dmitry A. Gordenin, *An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers*, Nature Genetics 45 (2013), no. 9, 970–976.

[200] P J Stephens, P S Tarpey, H Davies, P Van Loo, C Greenman, D C Wedge, S Nik-Zainal, S Martin, I Varela, G R Bignell, L R Yates, E Papaemmanuil, D Beare, A Butler, A Cheverton, J Gamble, J Hinton, M Jia, A Jayakumar, D Jones, C Latimer, K W Lau, S McLaren, D J McBride, A Menzies, L Mudie, K Raine, R Rad, M S Chapman, J Teague, D Easton, A Langerod, Consortium Oslo Breast Cancer, M T Lee, C Y Shen, B T Tee, B W Huimin, A Broeks, A C Vargas, G Turashvili, J Martens, A Fatima, P Miron, S F Chin, G Thomas, S Boyault, O Mariani, S R Lakhani, M van de Vijver, L van 't Veer, J Foekens, C Desmedt, C Sotiriou, A Tutt, C Caldas, J S Reis-Filho, S A Aparicio, A V Salomon, A L Borresen-Dale, A L Richardson, P J Campbell, P A Futreal, and M R Stratton, *The landscape of cancer genes and mutational processes in breast cancer*, Nature 486 (2012), no. 7403, 400–404.

[201] L M Cortez, A J L Brown, M A Dennis, C D Collins, A J L Brown, D Mitchell, T M Mertz, and S A Roberts, *APOBEC3A is a prominent cytidine deaminase in breast cancer*, PLoS Genet 15 (2019), no. 12, e1008545.

[202] Mia Petljak, Ludmil B. Alexandrov, Jonathan S. Brammeld, Stacey Price, David C. Wedge, Sebastian Grossmann, Kevin J. Dawson, Young Seok Ju, Francesco Iorio, Jose M.C. Tubio, Ching Chiek Koh, Ilias Georgakopoulos-Soares, Bernardo Rodríguez-Martín, Burçak Otlu, Sarah O'Meara, Adam P. Butler, Andrew Menzies, Shriram G. Bhosle, Keiran Raine, David R. Jones, Jon W. Teague, Kathryn Beal, Calli Latimer, Laura O'Neill, Jorge Zamora, Elizabeth Anderson, Nikita Patel, Mark Maddison, Bee Ling Ng, Jennifer Graham, Mathew J. Garnett, Ultan McDermott, Serena Nik-Zainal, Peter J. Campbell, and Michael R. Stratton, *Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis*, Cell 176 (2019), no. 6, 1282–1294.e20.

[203] Molly OhAinle, Julie A. Kerns, Melody M.H. Li, Harmit S. Malik, and Michael Emer-

man, *Antiretroelement Activity of APOBEC3H Was Lost Twice in Recent Human Evolution*, Cell Host and Microbe 4 (2008), no. 3, 249–259.

[204] X. Wang, A. Abudu, S. Son, Y. Dang, P. J. Venta, and Y.-H. H Zheng, *Analysis of Human APOBEC3H Haplotypes and Anti-Human Immunodeficiency Virus Type 1 Activity*, Journal of Virology 85 (2011), no. 7, 3142–3152.

[205] M M Li and M Emerman.

[206] Nicholas M. Chesarino and Michael Emerman, *Polymorphisms in human APOBEC3H differentially regulate ubiquitination and antiviral activity*, Viruses 12 (2020), no. 4.

[207] Jennifer A. Bohn, Keyur Thummar, Ashley York, Alice Raymond, W. Clay Brown, Paul D. Bieniasz, Theodora Hatziioannou, and Janet L. Smith, *APOBEC3H structure reveals an unusual mechanism of interaction with duplex RNA*, Nature Communications 8 (2017), no. 1, 1021.

[208] Jiang Gu, Qihan Chen, Xiao Xiao, Fumiaki Ito, Aaron Wolfe, and Xiaojiang S. Chen, *Biochemical Characterization of APOBEC3H Variants: Implications for Their HIV-1 Restriction Activity and mC Modification*, Journal of Molecular Biology 428 (2016), no. 23, 4626–4638.

[209] Pavel Silvestrov, Sarah J. Maier, Michelle Fang, and G. Andrés Cisneros, *DNArCdb: A database of cancer biomarkers in DNA repair genes that includes variants related to multiple cancer phenotypes*, DNA Repair 70 (2018), 10–17.

[210] Bernard W Stewart and Christopher P Wild, *Book Reviews World Cancer Report 2014*, vol. XV, World Health Organization, 2015.

[211] Alison C. MacKinnon, Jens Kopatz, and Tariq Sethi, *The molecular and cellular biology of lung cancer: Identifying novel therapeutic strategies*, British Medical Bulletin 95 (2010), no. 1, 47–61.

[212] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, J. Ferlay, A. Znaor, I. Soerjomataram, and F. Bray, *Global Cancer Observatory: Cancer Today*, 2018.

[213] Keith M. Kerr, *Pulmonary adenocarcinomas: Classification and reporting*, jan 2009, pp. 12–27.

[214] Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande, *OpenMM 7: Rapid development of high performance algorithms for molecular dynamics*, PLoS Computational Biology 13 (2017), no. 7, e1005659.

[215] Yuqing Feng, Robin P Love, Anjuman Ara, Tayyba T Baig, Madison B Adolph, and Linda Chelico, *Natural polymorphisms and Oligomerization of Human APOBEC3H contribute to single-stranded DNA scanning ability*, Journal of Biological Chemistry 290 (2015), no. 45, 27188–27203.

[216] Ramu Anandakrishnan, Boris Aguilar, and Alexey V. Onufriev, *"h++ 3.0: automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulation"*, Nucleic Acids Research 40 (2012), 537–541.

[217] Jesús A. Izaguirre, Chris R. Sweet, and Vijay S. Pande, *Multiscale dynamics of macromolecules using Normal Mode Langevin*, Pacific Symposium on Biocomputing 15 (2010), 240–251.

[218] Sachini U. Siriwardena, Kang Chen, and Ashok S. Bhagwat, *Functions and Malfunctions of Mammalian DNA-Cytosine Deaminases*, Chemical Reviews 116 (2016), no. 20, 12688–12710.

[219] Irene M. Ward and Junjie Chen, *Histone H2AX Is Phosphorylated in an ATR-dependent Manner in Response to Replicational Stress*, Journal of Biological Chemistry 276 (2001), no. 51, 47759–47762.

[220] Emmy P. Rogakou, Duane R. Pilch, Ann H. Orr, Vessela S. Ivanova, and William M. Bonner, *DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139*, Journal of Biological Chemistry 273 (1998), no. 10, 5858–5868.

[221] J M Kidd, T L Newman, E Tuzun, R Kaul, and E E Eichler.

[222] N K Duggal, H S Malik, and M Emerman.

[223] Cristina J. Wittkopp, Madison B. Adolph, Lily I. Wu, Linda Chelico, and Michael

Emerman, *A Single Nucleotide Polymorphism in Human APOBEC3C Enhances Restriction of Lentiviruses*, PLoS Pathogens 12 (2016), no. 10, e1005865.

[224] Yuqing Feng, Lai Wong, Michael Morse, Ioulia Rouzina, Mark C. Williams, and Linda Chelico, *RNA-Mediated Dimerization of the Human Deoxycytidine Deaminase APOBEC3H Influences Enzyme Activity and Interaction with Nucleic Acids*, Journal of Molecular Biology 430 (2018), no. 24, 4891–4907.

[225] Linda Chelico, Courtney Prochnow, Dorothy A Erie, Xiaojiang S Chen, and Myron F Goodman, *Structural model for deoxycytidine deamination mechanisms of the HIV-1 inactivation enzyme APOBEC3G.*, The Journal of biological chemistry 285 (2010), no. 21, 16195–205.

[226] David W Cescon, Benjamin Haibe-Kains, and Tak W Mak, *APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation*, Proceedings of the National Academy of Sciences of the United States of America 112 (2015), no. 9, 2841–2846.

[227] Katarzyna Klonowska, Wojciech Kluzniak, Bogna Rusak, Anna Jakubowska, Magdalena Ratajska, Natalia Krawczynska, Danuta Vasilevska, Karol Czubak, Marzena Wojciechowska, Cezary Cybulski, Jan Lubinski, and Piotr Kozlowski, *The 30 kb deletion in the APOBEC3 cluster decreases APOBEC3A and APOBEC3B expression and creates a transcriptionally active hybrid gene but does not associate with breast cancer in the European population*, Oncotarget 8 (2017), no. 44, 76357–76374.

[228] Jingjing Liu, Anieta M. Sieuwerts, Maxime P. Look, Michelle Van Der Vlugt-Daane, Marion E. Meijer-Van Gelder, John A. Foekens, Antoinette Hollestelle, and John W.M. Martens, *The 29.5 kb APOBEC3B deletion polymorphism is not associated with clinical outcome of breast cancer*, PLoS ONE 11 (2016), no. 8, e0161731.

[229] Wei Xiong Wen, Jaslyn Sian Siu Soo, Pui Yoke Kwan, Elaine Hong, Tsung Fei Khang, Shivaani Mariapun, Christine Shu Mei Lee, Siti Norhidayu Hasan, Pathmanathan Rajadurai, Cheng Har Yip, Nur Aishah Mohd Taib, and Soo Hwang Teo, *Germline APOBEC3B*

deletion is associated with breast cancer risk in an Asian multi-ethnic cohort and with immune cell presentation, Breast Cancer Research 18 (2016), no. 1, 56.

[230] Rémi Buisson, Adam Langenbucher, Danae Bowen, Eugene E Kwan, Cyril H Benes, Lee Zou, and Michael S Lawrence, *Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features*, Science 364 (2019), no. 6447.

[231] R. Barroso-Sousa, E. Jain, O. Cohen, D. Kim, J. Buendia-Buendia, E. Winer, N. Lin, S. M. Tolaney, and N. Wagle, *Prevalence and mutational determinants of high tumor mutation burden in breast cancer*, Annals of Oncology 31 (2020), no. 3, 387–394.

[232] Artur A. Serebrenik, Prokopios P. Argyris, Matthew C. Jarvis, William L. Brown, Martina Bazzaro, Rachel I. Vogel, Britt K. Erickson, Sun Hee Lee, Krista M. Goergen, Matthew J. Maurer, Ethan P. Heinzen, Ann L. Oberg, Yajue Huang, Xiaonan Hou, S. John Weroha, Scott H. Kaufmann, and Reuben S. Harris, *The DNA cytosine deaminase APOBEC3B is a molecular determinant of platinum responsiveness in clear cell ovarian cancer*, Clinical Cancer Research 26 (2020), no. 13, 3397–3407.

[233] Marcel Ooms, Bonnie Brayton, Michael Letko, Susan M. Maio, Christopher D. Pilcher, Frederick M. Hecht, Jason D. Barbour, and Viviana Simon, *HIV-1 Vif adaptation to human APOBEC3H haplotypes*, Cell Host and Microbe 14 (2013), no. 4, 411–421.

[234] Eric W. Refsland, Judd F. Hultquist, Elizabeth M. Luengas, Terumasa Ikeda, Nadine M. Shaban, Emily K. Law, William L. Brown, Cavan Reilly, Michael Emerman, and Reuben S. Harris, *Natural Polymorphisms in Human APOBEC3H and HIV-1 Vif Combine in Primary T Lymphocytes to Affect Viral G-to-A Mutation Levels and Infectivity*, PLoS Genetics 10 (2014), no. 11, e1004761.

[235] M B Adolph, A Ara, Y Feng, C J Wittkopp, M Emerman, J S Fraser, and L Chelico.

[236] Nisha K Duggal and Michael Emerman, *Evolutionary conflicts between viruses and restriction factors shape immunity*, 2012, pp. 687–695.

[237] E K Law, A M Sieuwerts, K LaPara, B Leonard, G J Starrett, A M Molan, N A Temiz, R I Vogel, M E Meijer-van Gelder, F C Sweep, P N Span, J A Foekens, J W Martens, D Yee, and R S Harris.

[238] David W. Cescon and Benjamin Haibe-Kains, *DNA replication stress: A source of APOBEC3B expression in breast cancer*, Genome Biology 17 (2016), no. 1, 202.

[239] Abby M. Green, Sébastien Landry, Konstantin Budagyan, Daphne C. Avgousti, Sophia Shalhout, Ashok S. Bhagwat, and Matthew D. Weitzman, *APOBEC3A damages the cellular genome during DNA replication*, Cell Cycle 15 (2016), no. 7, 998–1008.

[240] Amit Gaba, Ben Flath, and Linda Chelico, *Examination of the apobec3 barrier to cross species transmission of primate lentiviruses*, Viruses 13 (2021), no. 6.

[241] Reuben S. Harris, Judd F. Hultquist, and David T. Evans, *The restriction factors of human immunodeficiency virus*, Journal of Biological Chemistry 287 (2012), no. 49, 40875–40883.

[242] Lucie Etienne, Beatrice H. Hahn, Paul M. Sharp, Frederick A. Matsen, and Michael Emerman, *Gene loss and adaptation to hominids underlie the ancient origin of HIV-1*, Cell Host and Microbe 14 (2013), no. 1, 85–92.

[243] Keiya Uriu, Yusuke Kosugi, Jumpei Ito, and Kei Sato, *The battle between retroviruses and APOBEC3 genes: Its past and present*, Viruses 13 (2021), no. 1, 1–11.

[244] Adam Z. Cheng, Sofia N. Moraes, Nadine M. Shaban, Elisa Fanunza, Craig J. Bierle, Peter J. Southern, Wade A. Bresnahan, Stephen A. Rice, and Reuben S. Harris, *APOBECs and herpesviruses*, Viruses 13 (2021), no. 3, 1–14.

[245] J F Arias, T Koyama, M Kinomoto, and K Tokunaga.

[246] Krista A Delviks-frankenberry, Belete A Desimmie, and Vinay K Pathak, *Structural Insights into APOBEC3-Mediated Lentiviral Restriction*, Viruses 4 (2020).

[247] Wendy Kaichun Xu, Hyewon Byun, and Jaquelin P. Dudley, *The role of APOBECs in viral replication*, Microorganisms 8 (2020), no. 12, 1–46.

[248] Darja Pollpeter, Maddy Parsons, Andrew E. Sobala, Sashika Coxhead, Rupert D. Lang, Annie M. Bruns, Stelios Papaioannou, James M. McDonnell, Luis Apolonia, Jamil A. Chowdhury, Curt M. Horvath, and Michael H. Malim, *Deep sequencing of HIV-1 reverse transcripts reveals the multifaceted antiviral functions of APOBEC3G*, Nature Microbiology 3 (2018), no. 2, 220–233.

[249] Y Iwatani, D S Chan, F Wang, K S Maynard, W Sugiura, A M Gronenborn, I Rouzina, M C Williams, K Musier-Forsyth, and J G Levin.

[250] Yingxia Hu, Kirsten M. Knecht, Qi Shen, and Yong Xiong, *Multifaceted HIV-1 Vif interactions with human E3 ubiquitin ligase and APOBEC3s*, FEBS Journal 288 (2021), no. 11, 3407–3417.

[251] A Ara, R P Love, and L Chelico.

[252] C. Chaipan, J. L. Smith, W.-S. Hu, and V. K. Pathak, *APOBEC3G Restricts HIV-1 to a Greater Extent than APOBEC3F and APOBEC3DE in Human Primary CD4+ T Cells and Macrophages*, Journal of Virology 87 (2013), no. 1, 444–453 (eng).

[253] S R Richardson, I Narvaiza, R A Planegger, M D Weitzman, and J V Moran.

[254] Judd F. Hultquist, Joy A. Lengyel, Eric W. Refsland, Rebecca S. LaRue, Lela Lackey, William L. Brown, and Reuben S. Harris, *Human and Rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H Demonstrate a Conserved Capacity To Restrict Vif-Deficient HIV-1*, Journal of Virology 85 (2011), no. 21, 11220–11234.

[255] Rayhane Nchioua, Dorota Kmiec, Amit Gaba, Christina M. Stürzel, Tyson Follack, Stephen Patrick, Andrea Kirmaier, Welkin E. Johnson, Beatrice H. Hahn, Linda Chelico, and Frank Kirchhoff, *APOBEC3F Constitutes a Barrier to Successful Cross-Species Transmission of Simian Immunodeficiency Virus SIVsmm to Humans*, Journal of Virology 95 (2021), no. 17.

[256] Brett D. Anderson, Terumasa Ikeda, Seyed Arad Moghadasi, Amber St Martin, William L. Brown, and Reuben S. Harris, *Natural APOBEC3C variants can elicit differential HIV-1 restriction activity*, Retrovirology 15 (2018), no. 1, 1–11.

[257] Ananda Ayyappan Jaguva Vasudevan, Kannan Balakrishnan, Christoph G.W. Gertzen, Fanni Borvető, Zeli Zhang, Anucha Sangwiman, Ulrike Held, Caroline Küstermann, Sharmistha Banerjee, Gerald G. Schumann, Dieter Häussinger, Ignacio G. Bravo, Holger Gohlke, and Carsten Münk, *Loop 1 of APOBEC3C Regulates its Antiviral Activity against HIV-1*, Journal of Molecular Biology 432 (2020), no. 23, 6200–6227.

[258] Cesar A. Virgen and Theodora Hatziioannou, *Antiretroviral Activity and Vif Sensitivity of Rhesus Macaque APOBEC3 Proteins*, Journal of Virology 81 (2007), no. 24, 13932–13937.

[259] Zeli Zhang, Qinyong Gu, Ananda Ayyappan Jaguva Vasudevan, Manimehalai Jeyaraj, Stanislaw Schmidt, Jörg Zielonka, Mario Perković, Jens-Ove Heckel, Klaus Cichutek, Dieter Häussinger, Sander H. J. Smits, and Carsten Münk, *Vif Proteins from Diverse Human Immunodeficiency Virus/Simian Immunodeficiency Virus Lineages Have Distinct Binding Sites in A3C*, Journal of Virology 90 (2016), no. 22, 10193–10208.

[260] J. Myers, G. Grothaus, S. Narayanan, and A. Onufriev, *A simple clustering algorithm can be accurate enough for use in calculations of pks in macromolecules*, Proteins 63 (2006), 928–938.

[261] J.C. Gordon, J.B. Myers, T. Folta, V. Shoja, L.S. Heath, and A. Onufriev, *H++: a server for estimating pkas and adding missing hydrogens to macromolecules*, Nucleic Acids Research 33 (2005), 368–371.

[262] Vincent B. Chen, W. Bryan Arendall, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson, and David C. Richardson, *MolProbity: All-atom structure validation for macromolecular crystallography*, Acta Crystallographica Section D: Biological Crystallography 66 (2010), no. 1, 12–21.

[263] John Towns, Tim Cockerill, Maytal Dahan, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, and Gregory D Peterson, *Scientific cyberinfraStructure XSEDE: Accelerating Scientific Discovery*, Computing in Science and Engineering (2014), no. September/October, 62–74.

[264] Emmett Leddin, Cisneros Research Group, and G. Andres Cisneros, *Cisnerosresearch/amber-eda: First release*, `https://github.com/CisnerosResearch/AMBER-EDA`, 01 2020.

[265] Harold C. Smith, *RNA binding to APOBEC deaminases; Not simply a substrate for C to U editing*, RNA Biology 14 (2017), no. 9, 1153–1165.

[266] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt, *Direct-*

coupling analysis of residue coevolution captures native contacts across many protein families, Proceedings of the National Academy of Sciences of the United States of America 108 (2011), no. 49.

[267] Ivan Anishchenko, Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker, *Origins of coevolution between residues distant in protein 3D structures*, Proceedings of the National Academy of Sciences of the United States of America 114 (2017), no. 34, 9122–9127.

[268] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander, *Protein 3D structure computed from evolutionary sequence variation*, PLoS ONE 6 (2011), no. 12.

[269] Ricardo N. Dos Santos, Faruck Morcos, Biman Jana, Adriano D. Andricopulo, and José N. Onuchic, *Dimeric interactions and complex formation using direct coevolutionary couplings*, Scientific Reports 5 (2015), 1–10.

[270] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker, *Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information*, eLife 2014 (2014), no. 3, 1–21.

[271] Duccio Malinverni, Alfredo Jost Lopez, Paolo De Los Rios, Gerhard Hummer, and Alessandro Barducci, *Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and coevolutionary sequence analysis*, eLife 6 (2017), 1–21.

[272] Martin Werner, Vytautas Gapsys, and Bert L. De Groot, *One plus One Makes Three: Triangular Coupling of Correlated Amino Acid Mutations*, Journal of Physical Chemistry Letters 12 (2021), no. 12, 3195–3201.

[273] S Kitamura, H Ode, M Nakashima, M Imahashi, Y Naganawa, T Kurosawa, Y Yokomaku, T Yamane, N Watanabe, A Suzuki, W Sugiura, and Y Iwatani.

[274] 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 (2017), no. 1, 1–12.

[275] Justin J. King and Mani Larijani, *A novel regulator of activation-induced cytidine*

deaminase/APOBECs in immunity and cancer: Schrödinger's CATalytic pocket, Frontiers in Immunology 8 (2017), no. APR, 7–9.

[276] Q Yu, R Konig, S Pillai, K Chiles, M Kearney, S Palmer, D Richman, J M Coffin, and N R Landau.

[277] Reuben S. Harris and Jaquelin P. Dudley, *APOBECs and virus restriction*, Virology 479-480 (2015), 131–145.

[278] Y Feng and L Chelico.

[279] R P Love, H Xu, and L Chelico.

[280] 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, Rommie E. Amaro, Daniel A. Harki, Reuben S. Harris, and Hideki Aihara, *Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B*, Nature Structural and Molecular Biology 24 (2017), no. 2, 131–139.

[281] Madison B. Adolph, Robin P. Love, and Linda Chelico, *Biochemical Basis of APOBEC3 Deoxycytidine Deaminase Activity on Diverse DNA Substrates*, ACS Infectious Diseases 4 (2018), no. 3, 224–238.

[282] A Ara, R P Love, T B Follack, K A Ahmed, M B Adolph, and L Chelico.

[283] E I Garcia and M Emerman.

[284] Shuichi Hoshika, Nicole A. Leal, Myong Jung Kim, Myong Sang Kim, Nilesh B. Karalkar, Hyo Joong Kim, Alison M. Bates, Norman E. Watkins, Holly A. SantaLucia, Adam J. Meyer, Saurja DasGupta, Joseph A. Piccirilli, Andrew D. Ellington, John SantaLucia, Millie M. Georgiadis, and Steven A. Benner, *Hachimoji DNA and RNA: A genetic system with eight building blocks*, Science 363 (2019), no. 6429, 884–887.

[285] Noriko Saito-Tarashima and Noriaki Minakawa, *Unnatural base pairs for synthetic biology*, Chemical and Pharmaceutical Bulletin 66 (2018), no. 2, 132–138.

[286] Perry C.M. and Noble S., *Didanosine: An updated review of its use in HIV infection*, Drugs 58 (1999), no. 6, 1099–1135.

[287] Heawon Ann, Ki Hyon Kim, Hyun Young Choi, Hyun Ha Chang, Sang Hoon Han,

Kye Hyung Kim, Jin Soo Lee, Yeon Sook Kim, Kyung Hwa Park, Young Keun Kim, Jang Wook Sohn, Na Ra Yun, Chang Seop Lee, Young Wha Choi, Yil Seob Lee, and Shin Woo Kim, *Safety and efficacy of Ziagen (abacavir sulfate) in HIV-infected Korean patients*, Infection and Chemotherapy 49 (2017), no. 3, 205–212.

[288] Peter L Anderson and Joseph E Rower, *Zidovudine and Lamivudine for HIV Infection*, Clinical Medicine Reviews in Therapeutics 2 (2010), 115–127.

[289] Julie C. Adkins, David H. Peters, and Diana Faulds, *Zalcitabine: An Update of Its Pharmacodynamic and Pharmacokinetic Properties and Clinical Efficacy in the Management of HIV Infection*, Drugs 53 (1997), no. 6, 1054–1080.

[290] Seng Gee Lim, Tay Meng Ng, Nelson Kung, Zahary Krastev, Miroslava Volfova, Petr Husa, Samuel S. Lee, Sing Chan, Mitchell L. Shiffman, Mary Kay Washington, Amy Rigney, Jane Anderson, Elsa Mondou, Andrea Snow, Jeff Sorbel, Richard Guan, and Franck Rousseau, *A double-blind placebo-controlled study of emtricitabine in chronic hepatitis B*, Archives of Internal Medicine 166 (2006), no. 1, 49–56.

[291] Ting Tsung Chang, Ching Lung Lai, Seung Kew Yoon, Samuel S. Lee, Henrique Sergio M. Coelho, Flair Jose Carrilho, Fred Poordad, Waldemar Halota, Yves Horsmans, Naoky Tsai, Hui Zhang, Daniel J. Tenney, Ricardo Tamez, and Uchenna Iloeje, *Entecavir treatment for up to 5 years in patients with hepatitis b e antigen-positive chronic hepatitis B*, Hepatology 51 (2010), no. 2, 422–430.

[292] Deepak N. Amarapurkar, *Telbivudine: A new treatment for chronic hepatitis B*, World Journal of Gastroenterology 13 (2007), no. 46, 6150–6155.

[293] John H. Beigel, Kay M. Tomashek, Lori E. Dodd, Aneesh K. Mehta, Barry S. Zingman, Andre C. Kalil, Elizabeth Hohmann, Helen Y. Chu, Annie Luetkemeyer, Susan Kline, Diego Lopez de Castilla, Robert W. Finberg, Kerry Dierberg, Victor Tapson, Lanny Hsieh, Thomas F. Patterson, Roger Paredes, Daniel A. Sweeney, William R. Short, Giota Touloumi, David Chien Lye, Norio Ohmagari, Myoung-don Oh, Guillermo M. Ruiz-Palacios, Thomas Benfield, Gerd Fätkenheuer, Mark G. Kortepeter, Robert L. Atmar, C. Buddy Creech, Jens Lundgren, Abdel G. Babiker, Sarah Pett, James D. Neaton, Timo-

thy H. Burgess, Tyler Bonnett, Michelle Green, Mat Makowski, Anu Osinusi, Seema Nayak, and H. Clifford Lane, *Remdesivir for the Treatment of Covid-19 — Final Report*, New England Journal of Medicine 383 (2020), no. 19, 1813–1826.

[294] Saeid Malek Zadeh, Elahe K. Astani, Zhe Chong Wang, Kamal Adhikari, Rajesh Rattinam, and Tsung Lin Li, *Theoretical study of intermolecular interactions between critical residues of membrane protein MraYAAand promising antibiotic muraymycin D2*, ACS Omega 5 (2020), no. 36, 22739–22749.

[295] Jadd Shelton, Xiao Lu, Joseph A. Hollenbaugh, Jong Hyun Cho, Franck Amblard, and Raymond F. Schinazi, *Metabolism, Biochemical Actions, and Chemical Synthesis of Anticancer Nucleosides, Nucleotides, and Base Analogs*, Chemical Reviews 116 (2016), no. 23, 14379–14455.

[296] Xingbang Hu, Haoran Li, Lei Zhang, and Shijun Han, *Tautomerism of uracil and 5-bromouracil in a microcosmic environment with water and metal ions. What roles do metal ions play?*, Journal of Physical Chemistry B 111 (2007), no. 31, 9347–9354.

[297] E R Kaufman, *Replication of DNA containing 5-bromouracil can be mutagenic in Syrian hamster cells.*, Molecular and Cellular Biology 4 (1984), no. 11, 2449–2454.

[298] James N. Wilson and Eric T. Kool, *Fluorescent DNA base replacements: Reporters and sensors for biological systems*, Organic and Biomolecular Chemistry 4 (2006), no. 23, 4265–4274.

[299] L. Marcus Wilhelmsson, *Fluorescent nucleic acid base analogues*, Quarterly Reviews of Biophysics 43 (2010), no. 2, 159–183.

[300] Renatus W. Sinkeldam, Nicholas J. Greco, and Yitzhak Tor, *Fluorescent analogs of biomolecular building blocks: Design, properties, and applications*, Chemical Reviews 110 (2010), no. 5, 2579–2619.

[301] Suresh Kumar, Viswanathan Chinnusamy, and Trilochan Mohapatra, *Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond*, Frontiers in Genetics 9 (2018), no. December, 1–14.

[302] Fei Gao and Sanjoy K. Das, *Epigenetic regulations through DNA methylation and*

hydroxymethylation: Clues for early pregnancy in decidualization, Biomolecular Concepts 5 (2014), no. 2, 95–107.

[303] Pawel Siedlecki and Piotr Zielenkiewicz, *Mammalian DNA methyltransferases*, Acta Biochimica Polonica 53 (2006), no. 2, 245–256.

[304] Amy R. Vandiver, Adrian Idrizi, Lindsay Rizzardi, Andrew P. Feinberg, and Kasper D. Hansen, *DNA methylation is stable during replication and cell cycle arrest*, Scientific Reports 5 (2015), 1–8.

[305] Achim Breiling and Frank Lyko, *Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond*, Epigenetics and Chromatin 8 (2015), no. 1, 1–9.

[306] Ting Shi, Pingping Han, Chun You, and Yi Heng P.Job Zhang, *An in vitro synthetic biology platform for emerging industrial biomanufacturing: Bottom-up pathway design*, Synthetic and Systems Biotechnology 3 (2018), no. 3, 186–195.

[307] Maryke Fehlau, Felix Kaspar, Katja F. Hellendahl, Julia Schollmeyer, Peter Neubauer, and Anke Wagner, *Modular Enzymatic Cascade Synthesis of Nucleotides Using a (d)ATP Regeneration System*, Frontiers in Bioengineering and Biotechnology 8 (2020), no. August, 1–10.

[308] Hocek2019, *Enzymatic Synthesis of Base-Functionalized Nucleic Acids for Sensing, Cross-linking, and Modulation of Protein-DNA Binding and Transcription*, Accounts of Chemical Research 52 (2019), no. 6, 1730–1737.

[309] NC Brown, LW Dudycz, and GE Wright, *Rational design of substrate analogues targeted to selectively inhibit replication-specific dna polymerases*, Drugs under experimental and clinical research 12 (1986), no. 6-7, 555—564.

[310] Minchen Chien, Thomas K. Anderson, Steffen Jockusch, Chuanjuan Tao, Xiaoxu Li, Shiv Kumar, James J. Russo, Robert N. Kirchdoerfer, and Jingyue Ju, *Nucleotide Analogues as Inhibitors of SARS-CoV-2 Polymerase, a Key Drug Target for COVID-19*, Journal of Proteome Research 19 (2020), no. 11, 4690–4697.

[311] Uwe T Bornscheuer and Matthias Höhne, *Protein Engineering: Methods and Protocols [Methods in Molecular Biology, Vol. 1685]*, 1685 (2018), 350.

[312] John F. Darby, Masakazu Atobe, James D. Firth, Paul Bond, Gideon J. Davies, Peter O'Brien, and Roderick E. Hubbard, *Increase of enzyme activity through specific covalent modification with fragments*, Chemical Science 8 (2017), no. 11, 7772–7779.

[313] Stefan Lutz, *Beyond directed evolution-semi-rational protein engineering and design*, Current Opinion in Biotechnology 21 (2010), no. 6, 734–743.

[314] Martina Pavlova, Martin Klvana, Zbynek Prokop, Radka Chaloupkova, Pavel Banas, Michal Otyepka, Rebecca C. Wade, Masataka Tsuda, Yuji Nagata, and Jiri Damborsky, *Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate*, Nature Chemical Biology 5 (2009), no. 10, 727–733.

[315] Scott J. Novick, Nikki Dellas, Ravi Garcia, Charlene Ching, Abigail Bautista, David Homan, Oscar Alvizo, David Entwistle, Florian Kleinbeck, Thierry Schlama, and Thomas Ruch, *Engineering an amine transaminase for the efficient production of a chiral sacubitril precursor*, ACS Catalysis 11 (2021), no. 6, 3762–3770.

[316] Justin B Siegel, Alexandre Zanghellini, Helena M Lovick, Gert Kiss, Abigail R Lambert, Jennifer L. St.Clair, Jasmine L Gallaher, Donald Hilvert, Michael H Gelb, Barry L Stoddard, Kendall N Houk, Forrest E Michael, and David Baker, *Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction*, Science 329 (2010), no. 5989, 309–313.

[317] Rachel L. Weller and Scott R. Rajski, *Design, synthesis, and preliminary biological evaluation of a DNA methyltransferase-directed alkylating agent*, ChemBioChem 7 (2006), no. 2, 243–245.

[318] Hailong Chen, Zhilai Wang, Haibo Cai, and Changlin Zhou, *Progress in the microbial production of S-adenosyl-l-methionine*, World Journal of Microbiology and Biotechnology 32 (2016), no. 9, 1–8.

[319] Rafael A. Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, Hong Ji, James B. Potash, Sarven Sabunciyan, and Andrew P. Feinberg, *The human colon cancer*

methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores, Nature Genetics 41 (2009), no. 2, 178–186.

[320] Dan Levy, Alex J. Kuo, Yanqi Chang, Uwe Schaefer, Christopher Kitson, Peggie Cheung, Alexsandra Espejo, Barry M. Zee, Chih Long Liu, Stephanie Tangsombatvisit, Ruth I. Tennen, Andrew Y. Kuo, Song Tanjing, Regina Cheung, Katrin F. Chua, Paul J. Utz, Xiaobing Shi, Rab K. Prinjha, Kevin Lee, Benjamin A. Garcia, Mark T. Bedford, Alexander Tarakhovsky, Xiaodong Cheng, and Or Gozani, *Lysine methylation of the NF-κ B subunit RelA by SETD6 couples activity of the histone methyltransferase GLP at chromatin to tonic repression of NF-κ B signaling*, Nature Immunology 12 (2011), no. 1, 29–36.

[321] Xiaodong Cheng, *Structure and Function of DNA Methyltransferases*, Annu. Rev. Biophys. Biomol. Struct. 24 (1995), 293–318.

[322] Jochem Deen, Charlotte Vranken, Volker Leen, Robert K. Neely, Kris P.F. Janssen, and Johan Hofkens, *Methyltransferase-Directed Labeling of Biomolecules and its Applications*, Angewandte Chemie - International Edition 56 (2017), no. 19, 5182–5200.

[323] Paul N. Devine, Roger M. Howard, Rajesh Kumar, Matthew P. Thompson, Matthew D. Truppo, and Nicholas J. Turner, *Extending the application of biocatalysis to meet the challenges of drug development*, Nature Reviews Chemistry 2 (2018), no. 12, 409–421.

[324] Anant R. Kapdi and Yogesh S. Sanghvi, *Chapter 1 - the future of drug discovery: The importance of modified nucleosides, nucleotides, and oligonucleotides*, Palladium-Catalyzed Modification of Nucleosides, Nucleotides and Oligonucleotides (Anant R. Kapdi, Debabrata Maiti, and Yogesh S. Sanghvi, eds.), Elsevier, 2018, pp. 1–18.

[325] Marek Wojciechowski, Honorata Czapinska, and Matthias Bochtler, *CpG underrepresentation and the bacterial CpG-specific DNA methyltransferase M.Mpel*, Proceedings of the National Academy of Sciences of the United States of America 110 (2013), no. 1, 105–110.

[326] Tong Wang and Rahul M. Kohli, *Discovery of an unnatural dna modification derived from a natural secondary metabolite*, Cell Chemical Biology 28 (2021), no. 1, 97–104.e4.

[327] Enguerran Vanquelef, Sabrina Simon, Gaelle Marquant, Elodie Garcia, Geoffroy

Klimerak, Jean Charles Delepine, Piotr Cieplak, , and François-Yves Dupradeau, *R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments*, Nucleic Acids Research 39 (2011), no. suppl_2, W511–W517.

[328] F. Wang, J.-P. Becker, P. Cieplack, and F.-Y. Dupradeau, *R.E.D. python: Object oriented programming for AMBER force fields.*, 2013.

[329] François-Yves Dupradeau, Adrien Pigache, Thomas Zaffran, Corentin Savineau, Rodolphe Lelong, Nicolas Grivel, Dimitri Lelong, Wilfried Rosanski, and Piotr Cieplak, *The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building*, Phys. Chem. Chem. Phys. 12 (2010), 7821–7839.

[330] Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman, *A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model*, The Journal of Physical Chemistry 97 (1993), no. 40, 10269–10280.

[331] Daniel R. Roe and Thomas E. Cheatham, *PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data.*, Journal of chemical theory and computation 9 (2013), no. 7, 3084–95.

[332] Robert W. Zwanzig, *High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases*, Journal of Chemical Physics 22 (1954), no. 8, 1420–1426.

[333] G. Andrés Cisneros, Min Wang, Peter Silinski, Michael C. Fitzgerald, and Weitao Yang, *Theoretical and experimental determination on two substrates turned over by 4-oxalocrotonate tautomerase*, Journal of Physical Chemistry A 110 (2006), no. 2, 700–708.

[334] Carlo Adamo and Vincenzo Barone, *Toward reliable density functional methods without adjustable parameters: The PBE0 model*, Journal of Chemical Physics 110 (1999), no. 13, 6158–6170.

[335] Bianca Manta, Frank M. Raushel, and Fahmi Himo, *Reaction mechanism of zinc-dependent cytosine deaminase from escherichia coli: A quantum-chemical study*, Journal of Physical Chemistry B 118 (2014), no. 21, 5644–5652.

[336] 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 (yCD)*, Journal of Computational Chemistry 37 (2016), no. 13, 1163–1174.