Title

Probing nucleotide substrate selectivity during viral replication of SARS-CoV-2 RNA Dependent RNA Polymerase

Author

Romero, Moises Ernesto

Publication Date

2023

UNIVERSITY OF CALIFORNIA,
IRVINE

Probing nucleotide substrate selectivity during viral replication of SARS-CoV-2 RNA Dependent RNA Polymerase

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in Chemistry

by

Moises Ernesto Romero

Dissertation Committee:
Assistant Professor Jin Yu, Chair
Professor Ioan Andricioaei
Professor Gregory Weiss

2023

DEDICATION

Dedicado a mis padres Rosa y Jose Romero, sin sus sacrificios en esta vida nunca podría perseguir mis sueños.

TABLE OF CONTENTS

Page

LIST OF FIGURES **vii**

LIST OF TABLES **xxv**

ACKNOWLEDGEMENTS **xxvi**

VITA **xxix**

ABSTRACT OF THE DISSERTATION **xxxii**

1 Introduction **1**

1.1 The SARS-CoV-2 Genome and Viral Replication Machinery 1

1.2 The SARS-Cov-2 RNA Dependent RNA Polymerase (nsp12) 3

1.2.1 The RdRp active site 3

1.2.2 Elongation Cycle 5

1.3 Molecular Dynamics Simulations 9

1.3.1 Brief History of AMBER force field updates 10

1.3.2 Biased MD: Umbrella Sampling 11

2 ATP and Remdesivir: A Rigorous Examination of Initial Binding and Insertion States **15**

2.1 Introduction 15

2.2 Computational Details 18

2.2.1 Constructing Initial Binding & Insertion structures 18

2.2.2 Generating a Force Field for RDV-TP 19

2.2.3 Docking 22

2.2.4 Equilibrium Simulation Details 22

2.3 Results 24

2.4 Discussion 30

3 Thermodynamic Quantification of Nucleotide Insertion: A Free Energy Analysis of ATP and Remdesivir 32

3.1 Introduction 32

3.2 Determining the Reaction Coordinate and Calculating Free Energy 33

3.2.1 Selecting Reference Structures 34

3.2.2 Target MD and Umbrella Sampling 34

3.2.3 Hydrogen Bond Analysis around NTP 37

3.3 Free Energies of Insertion for ATP and RDV-TP 41

3.3.1 Insertion of ATP into the active site can be facilitated by base pairing with the stabilized template nt (+1) 42

3.3.2 RDV-TP initial stacking with the +1 template nt is more stabilized than the base pairing 46

3.3.3 RDV-TP insertion to the active site is facilitated by thermal fluctuations and absence of motif F-R553/R555 interaction with the polyphosphate 48

3.4 Discussion 50

3.5 Conclusions 55

4 dATP and GTP Vs ATP and RDV: A Comparative Study of Initial Binding and Insertion Through Equilibrium Ensemble Simulations 57

4.1 Introduction 57

4.2 Computational Details 60

4.2.1 Construction of Initial Binding and Insertion States of Noncognate: dATP and GTP 60

4.2.2 Simulation Parameters and Setup 60

4.3 Analysis 61

4.3.1 RMSD 61

4.3.2 Base Pair Geometry 62

4.3.3 Hydrogen Bond Occupancy 62

4.4 Equilibrium ensemble simulations: protein structural variations and distinctive dynamical responses to different NTPs upon initial binding and insertion 62

5 Nucleotide Selectivity: Comparative Assessment of Insertion Free Energy for dATP, GTP, ATP, and RDV 70

5.1 Introduction 70

5.2 Reaction Coordinate for GTP and dATP 73

5.3 Results 75

5.3.1 Constructing the PMFs for noncognate GTP/dATP from initial binding to insertion 76

5.3.2 Trapping noncognate dATP/GTP upon initial binding by persistent HB interactions from motif A/F/G, to NTP, template nt +1, or 3' end primer. 82

5.4 Discussion 87

5.4.1 Free energetics favor insertion of cognate NTPs but disfavor insertion of non-cognate NTPs 88

5.4.2 Key residues from conserved motifs detect and stabilize the ncNTP and its surroundings at the initial binding off-path state 90

5.5 Conclusions 92

6 Future Perspectives:Other Enhanced Sampling Methods and Improvement

of the RC **94**

6.1 Improving Sampling 94

6.1.1 Hamiltonian Replica Exchange 94

6.1.2 Metadynamics 96

6.2 Improving Reaction Coordinate 97

Bibliography **99**

Appendix **110**

LIST OF FIGURES

Page

1.1 The SARS-CoV-2 genome from the ∼30,000 kilobases in the +ssRNA. The ORF1a and ORF1b encode for the nonstructural proteins 1-16. Of key importance are those involved in replication and transcription shown at the *Bottom*. The nsp 12 (RNA dependent RNA polymerase) with cofactors nsp7 and nsp8 make up the core with an RNA duplex: template (blue) amd primer (red). Also shown are the Nsp13 helicase enzyme and Nsp14 exonuclease enzyme. . 2

1.2 The right hand structure of the pol domain from different viruses and the human mitochondria. With the subdomains colored: fingers (blue). palm (pink), and thumb (green).*Upper Left:* The SARS-Cov-2 RdRp pol subdomains and N-terminal domain in transparent gray (PDBid:7BV2). *Upper Right:* The Poliovirus RdRp which structurally similar to CoV-2 (PDBid:5f8j). *Lower Right:* The bacteriophage T7 viral DNA-directed RNA polymerase (PDBid:1s76). *Lower Left:* The DNA-directed RNA polymerase from the human mitochondria (model create from PDBid:4BOC). 4

1.3 The conserved motifs within the Pol domains of various polymerases are essential for their function. As depicted in the *upper panel*, Motifs A-G are structurally conserved across all viral RNA-dependent RNA polymerases (RdRps), and are associated with an RNA duplex (transparent red), magnesium ions (silver), and NTP and template strands (colored by atom name). In contrast, as shown in the *lower panel*, only Motifs A-D are present in DNA-directed RNA polymerases, which are associated with an RNA single strand (red), DNA duplex (transparent orange), magnesium ions (silver), and NTP and template strands (colored by atom name). Motif A (gray) and Motif C (green) contain conserved aspartic acid residues that play a critical role in catalysis. Motif B (orange) consists of residues that interact with the ribose sugar of the RNA. Motif D (pink) is speculated to contribute structural stability or facilitate the appropriate chemical environment for catalysis. Motif G (yellow) is involved in interactions with the template strand. The function of Motif E (light blue) remains elusive. Motif F (purple) is known to interact with the triphosphate moiety of the NTP and is hypothesized to be involved in fidelity control during replication. A structural element in RNA polymerase with a function analogous to Motif F is the O-helix, depicted in transparent purple. Structures are the same as used in Fig. 1.2. 6

1.4 This is a schematic representation of the elongation cycle for single nucleotide addition by the SARS-CoV-2 RdRp. Initial binding states or 'open' states are delineated by a green border, while the 'closed' or insertion states are marked in blue. Corresponding cryo-EM structures, identified by their respective PDB IDs, are depicted beneath their respective states and are shown in gray. . . . 7

1.5 The biophysical timescales with relevant motions for reference. Equilibrium molecular dynamics simulations (black line) are limited to the microsecond to sub-millisecond timescale. 12

2.1 SARS-CoV-2 RdRp elongation complex with an incorporated remdesivir (RDV) in the closed state (based on PDB:7BV2[11]). **A** The two main domains (N-terminus domain in grey, polymerase or pol domain in purple) of the RdRp along with the three cofactors (nsp8's in blue and nsp7 in green). **B** The pol domain consists three subdomains, the thumb (green), fingers (pink), and palm (blue). RNA (red) is shown along with incoming NTP and +1 template nt (red licorice). **C** Motifs A-G within the pol domain. 17

2.2 Comparing **A.** Adenosine Triphosphate (ATP) and **B.** Remdesivir Triphosphate (RDV-TP) heavy atom molecular structures. Hydrogens are omitted for a clear representation. Atoms colored in red highlight the differences in RDV from ATP. For partial charge calculation the RDV-TP is truncated at the O5' with the addition of a hydrogen H5T. H3T is the hydrogen atom bonded to the O3' oxygen (see Table. 2.1). 20

2.3 All-atom molecular dynamics simulation box. The size of the box on average: 15.7nm x 15.7 nm x 15.7 nm, containing an average of 382,000 atoms. 23

2.4 Docking of RDV-TP (or RTP) and ATP onto a modeled initial binding (active-site open) structure of SARS-CoV-2 RdRp (PDB: 7BTF). **A.** RDV-TP docking show comparatively stabilized docking structures (grouped into two). The palm, fingers, thumb subdomains are shown in red, blue, and green, and RNA in violet. The modeled RDV-TP (positioning from PDB: 7BV2) is shown in gray, and the docking structures of RDV-TP are shown with colored atoms. **B.** ATP docking shows diverse configurations and less stabilized configurations. The obtained docking energetics are listed on the right side for both systems (using AutoDock)[78] 24

2.5 Examining ATP binding configurations to the open form active site CoV-2 RdRp. **A.** The MD equilibration of the optimal docking complex of ATP to the CoV-2 RdRp structure. The equilibrated configurations were measured by RMSDs for both structural motifs (A-G) and ATP+template nt (uracil) with respect to the substrate insertion complex. Two dominant configurations of ATP have been identified, both of which are quite close to the insertion configuration (as our modeled open or initial binding complex of ATP to the RdRp, see Methods 2.1). **B.** The alignment of our modeled ATP bound open equilibrated form CoV-2 RdRp with that of the poliovirus (PV) RdRp, shown in two views for better visualization. 25

2.6 Modeled insertion structural complexes of SARS-CoV-2 for RDV-TP and ATP. Left and Center: The active site views with inserted ATP and RDV-TP shown at the end of equilibrium simulations for the insertion (A & C) and initial binding (B & D) states. Right: The open and closed RdRp structures aligned (E), with ATP initial binding and inserted shown, respectively. The CoV-2 RdRp is shown in comparison with previously studied RdRp from Poliovirus (PV) (F) (PDBs: 3ola and 3ol7).[16] 26

2.7 Expected hydrogen bond (HB) distance between ATP/RDV-TP and the +1 template Uracil (**A-E**). The distance pairs measured are the heavy atoms from the nucleotide triphosphate N6 with U:O4 and N1 and U:N3. The dashed black line indicate the cutoff (3.5 Å) for a HB. The NTPs in the insertion complexes (**B&E**) form significantly more stable HB than in in the initial binding forms. An RDV-TP in initial-binding forms a base stacking configuration with the template nt (**D**), in which hydrogen bonds are rarely formed. 27

2.8 Measured RMSD from equilibration simulations. (Left) Subdomains, RNA, and ATP. (Right) Cofactors (ns7 and nsp8), Nsp12 and N-Terminus domain. **A** ATP initial binding complex. **B** ATP insertion complex. The insertion complex appears to be more stable than the initial binding complex. 28

2.9 Measured RMSD from RTP (RDV-TP) equilibration simulations. (Left) Subdomains, RNA, and NTP. (Right) Cofactors (nsp and nsp8), Nsp12 and N-Terminus domain. **A** RDV-TP initial binding base pairing configuration. **B** RDV-TP initial binding stacking configuration. **C** RDV-TP insertion complex. The RDV-TP insertion complex appears to be more stable than the initial binding configurations. 29

3.1 Implementation of targeted molecular dynamics (TMD) simulations for constructing NTP insertion path to be utilized in the umbrella sampling simulations. (Left) The initial and final structures of each respective paths: backward (bottom) and forward (upper). With motifs A/D colored pink for the starting structure and green for the target structure. Representations are colored to compare with PV RdRp in.[13] (Right) The implementation of the TMD simulations forward and backward. Where TMD from open refers to starting from the initial binding complex (forward path), and TMD from closed is starting from the insertion complex (backward path). Structures are selected every 0.1 Å from the two paths until they meet in the middle along the reaction coordinate (RC). Such a constructed path is then used for the umbrella sampling simulation. 35

3.2 Conducting the umbrella sampling simulations for the NTP insertion. (Left) The schematics of the umbrella sampling simulation strategies (figure adapted from[100]). (Right) The overlap of simulated windows, where the RC is centered every 0.1Å, with the initial simulation structure taken from the forward/backward TMD paths (see Fig. 3.1). 37

3.3 Salt-bridge electrostatic interactions with ATP/RDV-TP triphosphates. Here we can identify the positively charged residues (Lys and Arg) which can form salt bridges with the negatively charged oxygen's along the polyphosphate. Distances are measured from the positive charge center (NZ nitrogen in Lys and CZ carbon in Arg) and the negative charge (O1G,O2G,O3G,O1B,O2B,O1A, and O2A in the NTP or O1P and O2P in the template backbone), if the charges are less than 5Å[104] a salt bridge is identified. 39

3.4 Schematics for base-stacking measurements. **A** The stacking is determined by measuring if the two base rings are parallel with reasonable overlapping. **B** The first $\cos\theta_1$ is calculated from atoms within the six numbered rings in the bases. **C** The second $\cos\theta_2$ is calculated via considering the center of mass (C.O.M.) of the two six numbered rings. Reasonable base stacking is formed when $\cos\theta_1 > 0.8$ (θ_1 close to zero or the two rings being parallel) and $\cos\theta_2 > 0.6$ (θ_2 not far from zero or the two rings overlap). Measurements shown are from the RDV-TP with base stacking (without force on template nt) insertion. 40

3.5 **A** The modeled insertion state structure of RDV-TP (grey transparent) aligned with that of the equilibrated insertion state one (in color). MgA is in coordination with residues D760 and D761. MgB is coordinated by the β and γ phosphate oxygens. **B-F** Distances measured between the center of geometry (C.O.G.) of the NTP and Mg^{2+} ions in the umbrella sampling simulations (from open to barrier and to the closed state). MgA is in coordination with the catalytic D760 and D761 residues as well as the 3' end primer backbone phosphate. MgB is in coordination with the phosphate group (β & γ phosphate oxygens).[19] MgC is unlikely to be involved in catalysis or product release as it stays comparatively far from the NTP. 41

3.6 Convergence plots of all PMFs constructed with bootstrapping error analysis for each set of data.[96;102] As more data are accumulated with the extended simulations, the PMF further converges. Early data collected is removed as for pre-equilibration, 20ns for ATP and 10ns for RDV-TP systems. **A&B** ATP PMF's with no force on template and with force on template, respectively. **C** RDV-TP base pairing with force on +1 template nt. **D** RDV-TP stacking with force on + 1 template nt. **E** RDV-TP stacking with no force on + 1 template. **D-E** Only 50ns of data from each window was needed to reach convergence. 43

3.7 ATP insertion from umbrella sampling MD simulation (without force on the +1 template nt in the RC). **A** PMF with barrier 5.0±0.3 kcal/mol and an initial binding stability of 4.8±0.3 kcal/mol. **B** Open conformation with ATP not forming hydrogen bonds with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows (see Fig. 3.3A for salt bridges). **D** Interaction snapshots from simulation windows: Two open states are shown due to the volatility of the open minima, ATP often flips out of plane from the +1 template base. As the barrier is crossed it begins to form consistent base pairing with the template. Dotted orange lines highlight essential HB interactions. 44

3.8 ATP insertion from umbrella sampling MD simulation (with force on the +1 template nt in the RC). **A** PMF with barrier of 2.6 ± 0.3 kcal/mol and an initial binding stability of 5.1±0.2 kcal/mol. **B** Open conformation of ATP forming hydrogen bonds with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3B for salt bridges). **D** Interaction snapshots from simulation windows: Two open states are shown due to the volatility of the open minima. Although ATP still occasionally flips out of plane, it more consistently forms HB with the +1 template base. Dotted orange lines highlight essential HB interactions. 45

3.9 RDV-TP insertion with the open state forming good base pairing from umbrella sampling MD simulation (with force on the +1 template nt in the RC). **A** PMF with barrier of 5.4 ± 0.3 kcal/mol and an initial binding stability of 2.6±0.3 kcal/mol. **B** Open conformation of RDV-TP forming hydrogen bonds with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3C for salt bridges). **D** Interaction snapshots from simulation windows: Throughout the open state stable HB form with the RDV and +1 template base. Dotted orange lines highlight essential HB interactions. . 46

3.10 RDV-TP insertion with the open state forming base stacking with the +1 template base from umbrella sampling MD simulation (with force on the +1 template nt in the RC). **A** PMF with barrier of 5.2 ± 0.3 kcal/mol and an initial binding stability of 2.6±0.2 kcal/mol. **B** Open conformation of RDV-TP forming base stacking with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3D for salt bridges). **D** Interaction snapshots from simulation windows: Throughout the open state base stacking forms resulting in a more stable minima. Dotted orange lines highlight essential HB interactions. 48

3.11 RDV-TP insertion with the open state forming base stacking with the +1 template base from umbrella sampling MD simulation (with no force on the +1 template nt in the RC). **A** PMF with barrier of 1.5 ± 0.2 kcal/mol and an initial binding stability of 2.7±0.1 kcal/mol. **B** Open conformation of RDV-TP forming base stacking with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3E for salt bridges). **D** Interaction snapshots from simulation windows: Throughout the open state base stacking forms resulting in a more stable minima. Dotted orange lines highlight essential HB interactions. 49

3.12 Hydrogen bond analyses around the 3'-end RNA primer nt (uracil base). (Upper) S759 forms frequent HBs with the 3'-end primer into both RDV and ATP insertion states, but only does that for ATP initial binding but not RDV-TP initial binding systems. R555 forms stable HB with 3'-end primer O4 oxygen in the RDV stacking initial binding state but no stable HB for RDV base pairing. (Lower) Open and closed states for RDV-TP from stacking insertion (no force on the template) showing the R555 and S759 HBs on the 3' end primer, respectively. 52

4.1 The root-mean-square displacements (RMSDs) of RdRp structural subdomains (backbone atoms), RNA (phosphate backbone), and NTP (heavy atoms) measured from ensemble MD simulations. The RMSDs are shown for ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) upon initial nucleotide binding (active site open; left) and insertion (active site closed; right) states. The subdomains are shown in different colors: fingers (blue). Palm (pink), thumb (green), RNA (red), and NTP (black). 63

4.2 The root-mean-square displacements (RMSDs) of RdRp structural motifs (backbone atoms) measured from ensemble MD simulations. The motif RMSDs are shown for ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) upon initial nucleotide binding (active site open; left) and insertion (active site closed; right) states. The seven key motif RMSD are displayed in different colors: motif A (gray), Motif B (orange), Motif C (green), Motif D (Pink), motif E (light blue), motif F (purple), and motif G (dark blue). Structural representations of those motifs along with NTP, uracil template nucleotide, and two catalytic MG ions are shown for each simulation system. The dotted black line indicates the reference group of motifs (B & C) for systems of inserted ATP/RDV-TP. 64

4.3 NTP-template association geometries from ensemble equilibrium simulations. The geometric measures (see **Methods**) between the uracil template +1 nucleotide (nt) and individual incoming NTP are displayed, for ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right), upon initial binding (left) and insertion (right) for each NTP species. Licorice representations of the NTP and template nt show the dominant geometries for each simulation system. Dotted lines indicate the hydrogen bonds for the standard Watson-Crick (WC) or wobble base (WB) pairing. 66

4.4 NTP and 3'-end primer association geometries sampled from equilibrium ensemble simulations. The geometric measures (see Methods) between the 3'-end primer and individual incoming NTP: ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) are demonstrated, upon initial binding (left) and insertion (right) for each NTP species. Licorice representations of the NTP and 3'-end primer show the dominant geometries for each simulation system. Distance is measured by the center of mass between the bases. Base plane angle is measured using the C1'-C2-C5 (3' end primer) and C1'-C7-C5 (NTP). 67

4.5 The hydrogen bond occupancy of the equilibrium ensembles for each NTP: ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) are demonstrated. Each unique interaction >10% population is considered. Two color-code sets are used: protein-NTP interactions use brown (polyphosphate), red (sugar), and blue (base), and for template-nt / 3' end primer-NTP interactions use light brown (polyphosphate), light red (sugar), and light blue (base). 68

4.6 The hydrogen bond occupancy of the equilibrium ensembles for protein-template nt Uracil (purple) and protein-3' end primer (pink) for each NTP simulated: ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) are demonstrated, upon initial binding (left) and insertion (right) for each NTP species. Each unique interaction >10% population is considered. . 69

5.1 The constructed structural models of the SARS-CoV-2 RdRp elongation complex in its initial NTP substrate binding (open) and then inserted (closed) states. Top left: The simulated elongation complex is depicted with the Nsp12 pol domain (purple) + N-terminal domain (gray), two cofactor nsp8 (blue) and nsp7 (green), RNA (red) and an NTP in the active site. Top center: The three major subdomains within the pol domain fingers in blue, palm in pink, and thumb in green. The RNA is also shown in red as well as the NTP shown in space filling spheres. Top right: The subdomains RMSD for initial binding and insertion states for ATP highlighting the subtle fingers and palm coming together. See Fig. 4.1 for the rest of the NTP's. Bottom Left: The modeled and simulation equilibrated ATP is shown as bound initially to an open active site, with the seven protein motifs highlighted in color. The boxes to the right show the initially bound RDV-TP, dATP, and GTP that were also modeled and equilibrated from the simulations. Bottom Right: The modeled and equilibrated ATP is shown in the insertion or the active site closed state, with the seven structures motifs shown as well, and the boxes to the right displaying the modeled and equilibrated insertion configurations of RDV-TP, dATP, and GTP to the closed active site. 72

5.2 The potentials of mean force (PMFs) calculated for various NTP from initial binding (active site open) to the insertion (closed) state via umbrella sampling simulations. The difference of RMSDs with respect to open and closed reference structures $\delta RMSD = RMSD(X, X_{\text{Open ref}}) - RMSD(X, X_{\text{Closed ref}})$, was used as the reaction coordinate in the PMF construction. The upper left panel shows the PMFs for GTP, with (dark green) and without (light green) force on the template +1 nucleotide. In both cases, PMFs are shown in comparison with the PMFs obtained for cognate ATP (blue) and drug RDV-TP (pink). The upper right panel shows the PMFs for dATP, with (dark purple) and without (magenta) force on the template +1 nucleotide. The lower left and lower right panels display convergence plots of PMFs for GTP and dATP, respectively, in current umbrella sampling simulations, without force implemented to the template +1 nucleotide. 75

5.3 The potentials of mean force (PMFs) calculated for NTP from initial binding (active site open) to the insertion (closed) state via umbrella sampling simulations. The difference of RMSDs with respect to open and closed reference structures $RMSD(X, X_{\text{Open Ref}}) - RMSD(X, X_{\text{Closed Ref}})$,[50] was used as the reaction coordinate in the PMF construction (see Methods). Top left shows the PMF for dATP (magenta) and top right for that of GTP (green), both in comparison with the PMFs obtained for cognate ATP (blue) and analog RDV-TP (pink).[50;145] Smoothing has been applied to each PMF for clarity (see original PMFs and converging tests in Fig. 5.2). Note that the placement of the PMF of GTP relative to that of ATP/RDV-TP is according to the alchemical calculation in the closed state[145] while the placement of the PMF of dATP relative to that of ATP/RDV-TP is still uncertain, while the relative binding free energy at the closed state, denoted as δg, is estimated between 2-7 kcal/mol (see text). Bottom left shows a summary of the different PMF profiles between cognate and noncognate insertions. Bottom right insertion free energy and barrier heights for each PMF shown on the top plots. 77

5.4 Results from steered MD (SMD) pulling GTP from the insertion (active-site closed) state well towards the initial binding (open) state at a rate of 1 $\mathring{A}$/ns (force constant 2.4 $\frac{kcal}{mol\mathring{A}^2}$). The reaction coordinate (RC) in pulling simulations, defined as the distance between the GTP and the active center (the center of mass of all C within 10 Å of the 3' RNA primer), is shown on the left panels. Instantaneous force applied in the SMD simulations is shown on the right panels. Raw data values, 100 window smoothed curves, and 10 window smoothed curves are drawn. ±1 standard deviation of the RCs in the open and closed wells from umbrella sampling are included (as the gray and pink bars on the left panels). Trial simulation 01 was run to a total of 550 ns. The other two trials 02 and 03 were run to a total of 300 ns. 79

5.5 Results from steered MD (SMD) pulling dATP from the insertion (active-site closed) state well towards the initial binding (open) state at a rate of 1 $\mathring{A}$/ns (force constant 2.4 $\frac{kcal}{mol\mathring{A}^2}$). The reaction coordinate (RC) in pulling simulations, defined as the distance between the dATP and the active center (the center of mass of all C within 10 Å of the 3' RNA primer), is shown on the left panels. Instantaneous force applied in the SMD simulations is shown on the right panels. Raw data values, 100 window smoothed curves, and 10 window smoothed curves are drawn. ±1 standard deviation of the RCs in the open and closed wells from umbrella sampling are included (as the gray and pink bars on the left panels). All trial simulations were run to a total of 300 ns. 80

5.6 The NTP-template association geometry distributions obtained from the umbrella sampling simulations (for PMF calculations) in comparison with that from ensemble equilibrium simulations for various NTP species. A kernel density estimate has been used to visualize the data. Each simulation system open to closed, for ATP, RDV-TP, dATP and GTP. The equilibrium ensemble distribution is shown (blue) along with that obtained from the umbrella sampling (w/ force on template in orange; w/out force on template in green). The black dot indicates the reference state used to generate the initial paths for the umbrella sampling, and grey dot the reference state used in the alchemical calculations[145] 82

5.7 Summary of hydrogen bonding (HB) interactions that stabilize non-cognate GTP/dATP or surroundings upon initial binding from umbrella sampling simulations. Four interacting partners are considered: protein, incoming NTP (ATP, RDV-TP, dATP, and GTP), uracil template nucleotide +1, and the 3'-end primer. *Center:* The HBs with occupancy >20% in the simulations were identified among these four interaction partners. The relative HB occupancy levels are shown for the initial binding dATP and GTP (along with RDV-TP) with respect to that of cognate ATP as reference (with bar; see **SI Figure S10** for full HB statistics for all the simulation systems). Notable HB interactions that stabilize the non-cognate GTP and dATP initial binding systems are particularly denoted, respectively. Schematic and cartoon of key motifs stabilizing GTP (*left*) and dATP (*right*), along with template and 3'-end primer. 83

5.8 The hydrogen bond (HB) occupancy in umbrella sampling trajectories representing the initial binding (open) state minima for each PMF. The top row shows the HB for each NTP from protein/template +1 nt/3'-end primer. The bottom row shows the protein HB on the template +1 nt/3'-end primer. For each NTP (ATP, RDV-TP, dATP, and GTP), only unique interactions with a population greater than 10% are considered. The same color code used in Figure S2 and S3 is used to represent the different types of interactions. . . 84

5.9 Comparing key interactions that stabilize the non-cognate GTP (and surroundings) and drug analogue RDV-TP in the initial binding system. The conserved protein motifs are shown in cartoon representation while interacting residues from these motifs are shown in licorice using the same color. A/C: Incoming GTP/RDV-TP and template nucleotide are shown in licorice colored by atom name. Hydrogen bonds (HBs) formed between GTP (or RDV-TP) and template nucleotide uracil/protein/3' end RNA primer. Orange circle highlights the essential interactions involved in 'trapping' non-cognate GTP phosphate in the initially bound state (A), which are absent for RDV-TP phosphate (C). B/D: Incoming GTP/RDV-TP is shown in transparent representation for clarity, 3'-end primer and template nucleotide are shown in licorice colored by atom name. HBs formed between the protein and template for stabilization, without proper WB base pairing for GTP-template (B) and with template base stacking in RDV-TP (D). For detailed HB occupancy plots see Fig. 5.8 115

5.10 Comparing the key interactions that stabilize the non-cognate dATP (and surroundings) and cognate ATP in the initial binding system. The conserved protein motifs are shown in cartoon representation while interacting residues from these motifs are shown in licorice using the same color. A/C: Incoming dATP/ATP and template nucleotide are shown in licorice colored by atom name. Hydrogen bonds (HBs) formed between dATP (or ATP) and template nucleotide uracil/protein/3'-end RNA primer. Orange circle highlights the strongest interactions involved in 'trapping' non-cognate dATP in the initially bound state. B/D: Incoming dATP/ATP is shown in transparent representation for clarity, 3'-end primer and template nucleotide are shown in licorice colored by atom name. HBs formed between the protein and template nucleotide uracil / 3'-end RNA primer for stabilization, in the absence of proper WC base pairing in dATP (B). HBs formed between the protein and template nucleotide uracil / 3'-end RNA primer for stabilization, with proper WC base pairing in ATP (D). For detailed HB occupancy plots see Fig. 5.8. 116

LIST OF TABLES

Page

2.1 Summary of Partial Charges used for the RDV-TP force field compared with ATP. Charges are separated by section of the NTP. 21

3.1 Summary of target MD and Umbrella Sampling parameters. The force constant used from the TMD simulations were carried over and used for the respective umbrella sampling simulations. Large force constants were used for the ATP simulations and smaller ones for RDV-TP simulations. Where the () in RDV-TP or RTP systems indicate the initial binding structure (open for the active site open state). 36

4.1 Average Motif RMSD from the initial binding and insertion equilibrium ensemble simulations using the closed state minimized structure as reference. Units are in Angstrom. 65

5.1 Umbrella Sampling Parameters: force constant k for each path and total number of windows used 74

ACKNOWLEDGMENTS

First and foremost, I want to thank my advisors Jin Yu and Ioan Andricioaei for their support and knowledge imparted to me over the course of my PhD. I am deeply grateful to Ioan, whose scientific creativity and introduction to biophysics research provided me with a crucial foundation as a scientist. Simultaneously, I would like to express my sincere gratitude to Jin. Her generosity in providing me a place in her lab and her unmatched passion for understanding protein-nucleic acid interactions has enriched my learning experience.

I would also like to thank my lab mates throughout the years from both the Andricioaei and Yu labs. From the Andricioaei lab: Jim McSally, Anupam Chatterjee, Dhiman Ray, and Trevor Gokey. In particular, I would like to thank Trevor whose discussions over coffee taught me so many meticulous details about MD. From the Yu Lab: Carmen Al Masari and Shannon J McElhenney, our discussions of our projects, methods, bugs, and lunches at Southern Spice have been some of the most enjoyable parts of my PhD.

Other mentors and people at UCI who have helped me establish myself as a scientist and educator: Doug Tobias, Bob Pelayo, and Ramesh D. Arasasingham. Your passion for education and encouragement all these years has motivated me to follow a path in the academic field.

To Phong Luong who always had the best advice and helped me navigate many difficult situations.

UCI has helped advance my career with many incredible programs. I am especially grateful for the opportunity to participate in DECADE and the CSULB pre-professor program. Thanks to my mentor from the CSULB pre-professor program: Joshua Cotter whom taught me so much about the behind the scenes of being a professor and provided me with invaluable career advice.

To the people that helped me survive my first years at UCI:Jessica Kelz, Wyeth Gibson, Wilfred Russel, Alan Robledo, Victor Duran, Shane W. Flynn, and the Santa Ana taco truck your friendship and conversations were an invaluable source of support. Thank you Shane for encouraging me to learn python and providing guidance and mentorship from the day we met.

Dr. Kimberley Cousins, my undergraduate advisor who I worked with on my first ever research project. I never thought a summer extracting vitamin C from bell peppers would eventually lead to me pursuing a full PhD. The foundation and encouragement from my first year of undergraduate to even now is one of the main reasons I have succeeded.

The UCI SACNAS chapter and all the incredible people I met from other departments has been a huge source of inspiration. All other friends and colleagues whom have inspired me and helped me throughout my degree: Heriberto Flores Zuleta, Rakia Dhaoui, Alissa Matus, and Ali Younis. Everyone here and more I have been so lucky to work with you on projects to make STEM and graduate school a more inclusive environment. I truly believe all of you represent the future for science. In particular, Ali Younis you have been a great source of motivation and support while we have only just met I feel like we have always known each other, mashallah.

To Jordan Quintero and David Bradley Price, my friends from the I.E. thank you for always reminding me of my roots.

To my parents whom when I told I was going to keep going to school for another half-decade simply supported my choice. To my siblings Miguel and Monica, thank you for keeping our parents company while I have lived away all these years.

A special thanks to my wonderful partner, Melissa Blacketer, whose patience, love, and support has been a source of peace during the toughest times. I am incredibly lucky to have you in my life.

Finally, I humbly acknowledge financial support from Graduate Opportunity Fellowship, Chemical and Structural Biology Training Fellowship, National Science Foundation (NSF) Award #2028935, Department of Chemistry DEI Fellowship. None of the costly simulations would have been possible without the use of super computing resources from Oak Ridge National Laboratory and UCI's HPC3.

Portions of Chapter 1 of this dissertation is a reprint of the material as it appears in "Dissecting nucleotide selectivity in viral RNA polymerases" *Computational and Structural Biotechnology Journal* 2021 19 p. 3339-3348., used with permission from Elsevier. The co-authors listed in this publication are Chunhong Long, Daniel La Rocco, Jin Yu.

Chapter 2 & 3 of this dissertation is a reprint of the material as it appears in "Probing remdesivir nucleotide analogue insertion to SARS-CoV-2 RNA dependent RNA polymerase in viral replication" *Molecular Systems Design and Engineering* 2021 6 p.888-902., used with permission from Royal Society of Chemistry. The co-authors listed in this publication are Chunhong Long, Daniel La Rocco, Anusha Mysore Keerthi, Dajun Xu, Jin Yu.

VITA

MOISES ERNESTO ROMERO

EDUCATION:

- Doctor of Philosophy in Chemistry, University of California, Irvine (2023)
- Master of Science in Chemistry, University of California, Irvine (2020)
- Bachelor of Science in Chemistry (Honors), California State University San Bernardino (2017)

POSITIONS:

- Graduate Researcher, Dept. of Chemistry, University of California, Irvine.
  Research Advisor: Jin Yu / Ioan Andricioaei (2017-2023)
- Undergraduate Researcher, Dept. of Chemistry, California State University San Bernardino.
  Research Advisor: Kimberley Cousins (2015-2017)

RESEARCH PUBLICATIONS:

- **Moises Ernesto Romero**, Shannon McElhenney, Jin Yu, Trapping non-cognate nucleotide upon initial binding for replication fidelity control in SARS-CoV-2 RNA dependent RNA polymerase. *Submitted* **2023**
- Chunhong Long, **Moises Ernesto Romero**, Liqiang Dai, Jin Yu, Energetic vs entropic stabilization between Remdesivir analogue and cognate ATP upon binding and insertion into active site of SARS-CoV-2 RNA dependent RNA polymerase. *Physical Chemistry Chemical Physics* **2023**
- **Moises Ernesto Romero**, Chunhong Long, Daniel La Rocco, Anusha Mysore Keerthi, Dajun Xu, Jin Yu, "Probing remdesivir nucleotide analogue insertion to SARS-CoV-2 RNA dependent RNA polymerase in viralreplication" *Molecular Systems Design and Engineering* **2021**
- Chunhong Long, **Moises Ernesto Romero**, Daniel La Rocco, Jin Yu, "Dissecting nucleotide selectivity in viral RNA polymerases" *Computational and Structural Biotechnology Journal* **2021** 19 p. 3339-3348.

PRESENTATIONS (select):

- **Chemistry Seminar at CSUSB**. Moises Ernesto Romero, Jin Yu "Nucleotide Selectivity of the SARS-CoV-2 RNA dependent RNA polymerase". 2023. Oral Presentation.

- **Biophysical Society Meeting**. Moises Ernesto Romero, Jin Yu "Probing nucleotide and remdesivir analogue insertion to SARS-CoV-2 RNA polymerase in viral replication". 2022. Poster Presentation.

- **UCI School of Physical Sciences**. Moises Ernesto Romero, Ali Younis, Heriberto Flores "Anti-racism in the Physical Sciences". 2020. Oral Presentation.

AWARDS AND HONORS (select):

- **CSU Long Beach Pre-Professor Fellowship**
  University of California, Irvine & CSU Long Beach (2022)
- **BPS 2022 Travel Award**
  Biophysical Society Annual Meeting Travel Award (2022)
- **Department of Chemistry DEI Fellow**
  University of California, Irvine (2020)
- **Poster Award**
  Center for Multiscale Cell Fate Research Poster Award
  University of California, Irvine (2020)
- **NSF Honorable Mention**
  University of California, Irvine (2019)
- **Chemical and Structural Biology Training Fellowship**
  University of California, Irvine (2018)
- **Graduate Opportunity Fellowship**
  Graduate Division
  University of California, Irvine (2017)

MENTORSHIP:

- **Chem 180: Undergraduate Research** 2021-2022
  Worked with 5 UCI undergraduate students developing simulations and force field parameters for nucleotide analogue drugs to be used in protein simulations.

- **Undergraduate Internship** 2020-2021
  Worked with a UCB undergraduate during the summer developing a protocol to calculate small molecule parameters for simulations. Led to two publications, student is now a graduate student at UCI.

- **UCI Math BioU and Math ExpLR** 2020
  Worked with a UCI undergraduate student as well as two high school students on a research project calculating solvents effects on protein-drug complex for the summer program.

TEACHING EXPERIENCE:

- **Teaching Assistant** 2018,2019,2020,2023
  University of California, Irvine
  Professor: Ramesh D. Arasasingham
  Course: General Chemistry (Undergraduate Level, Ch 1A,1B,1C)

- **Teaching Assistant** 2018,2019,2022
  University of California, Irvine
  Professor: Kimberley Edwards
  Course: General Chemistry Laboratory (Undergraduate Level, Ch 1LD & Ch 1LC)

- **Teaching Assistant** 2017
  University of California, Irvine
  Professor: John C. Hemminger
  Course: Chemical Thermodynamics, Kinetics, and Dynamics (Undergraduate Level, Ch.132A)

SERVICE:

- **Society for the Advancement of Chicanos/Hispanics and Native Americans in Science**
  Secretary 2021-2022
  President 2020-2021

- **Department of Chemistry DECADE Student Representative** 2019-2022

- **S-Stem Alumni Panel**
  Participated in a career panel at CSU San Bernardino for new transfer students. 2019

- **Southern California Undergraduate Research Symposium**
  Poster judge for undergraduate research projects. 2018

ABSTRACT OF THE DISSERTATION

Probing nucleotide substrate selectivity during viral replication of SARS-CoV-2 RNA Dependent RNA Polymerase

by

Moises Ernesto Romero

Doctor of Philosophy in Chemistry

University of California, Irvine, 2023

Assistant Professor Jin Yu, Chair

The RNA dependent RNA polymerase (RdRp) in SARS-CoV-2, the virus responsible for the COVID-19 pandemic, is a highly conserved enzyme responsible for viral genome replication/transcription. While there are many SARS-CoV-2 variants, the RdRp protein has remained relatively conserved, making it an attractive target for antiviral drugs. This dissertation investigates the nucleotide addition cycle (NAC) and nucleotide selectivity during the viral RdRp elongation, focusing on an early stage of the cycle from initial nucleotide substrate binding (enzyme active site open) to rate-limiting insertion states (active site closed). This is in contrast to common computational or modeling works which examine a generic one-step substrate binding process. The interactions of the RdRp with representative incoming nucleoside triphosphates (NTPs) are studied: cognate ATP, RDV-TP (a drug analogue to ATP), non-cognates dATP and GTP, according to RNA template uracil. Ensemble equilibrium all-atom molecular dynamics (MD) simulations have been employed to explore the configuration space of each NTP in two kinetic states (open and closed). Due to the expected millisecond conformational change (from the open to closed) accompanying nucleotide insertion and selection, enhanced sampling methods have been conducted to calculate the free energy profiles or potentials of mean force (PMFs) of the NTP's. The analyses reveal a marked difference in the stabilization of cognate ATP and the RDV-TP analog versus non-cognate dATP and GTP. Upon initial binding and subsequent insertion, ATP and RDV-TP show marginal free energy barriers, whereas dATP and GTP show substantial stabilization upon initial binding followed by notably high barriers for insertion into the active site. This pattern suggests an intrinsic mechanism of nucleotide selectivity in RdRp that rejects non-cognate NTPs. Specifically, ATP and RDV-TP, which are selected for incorporation, are favored in the closed or insertion state, while non-cognate dATP and GTP appear trapped *off-path* in the open or initial binding state. These mechanisms are facilitated by conserved structural motifs in the RdRp's palm and fingers subdomain. Interestingly, the RDV-TP analog exhibits base stacking with the template Uracil upon initial binding, contrasting with the Watson-Crick base pairing seen between cognate ATP and the template. Moreover, our study shows that while RDV-TP drug analog stabilization from initial binding to insertion is primarily energetically driven, the stabilization of natural cognate ATP is also contributed entropically. This dissertation offers physical insights into the nucleotide insertion and selection processes of SARS-CoV-2 RdRp prior to catalysis and can support the development of antiviral drugs targeting viral RdRps.

Chapter 1

Introduction

Since the emergence of SARS-CoV-2, the virus responsible for the COVID-19 pandemic, there has been a monumental effort from the scientific community to understand it. Central to this understanding is the virus's replication machinery, which at its core contains the nonstructural protein 12 (nsp12). Nsp12 is an RNA-dependent RNA polymerase (RdRp) that is vital for the replication and transcription of the virus's genome. In this chapter, I will begin by providing a brief overview of the SARS-CoV-2 virus, including its genome and the enzymes involved in its lifecycle. Subsequently, I will narrow the focus to the RdRp, discussing the current state of knowledge regarding nsp12 in CoV-2, as well as drawing parallels with RdRps in other viruses. Finally, the chapter will conclude with a brief discussion on the computational methods employed to study the RdRp of CoV-2.

1.1 The SARS-CoV-2 Genome and Viral Replication Machinery

SARS-CoV-2 (CoV-2) is a virus that belongs to the Coronaviridae family[1] and possesses a positive-sense single-stranded RNA (+ssRNA) genome, which is notably one of the largest among RNA viruses, measuring approximately 30 kilobases (kb) in length.[2] Upon infecting host cells, the +ssRNA undergoes transcription to synthesize viral proteins. To effectively replicate and transcribe its extensive genome, CoV-2 employs an array of nonstructural proteins (nsps), many of which synergize as part of the viral replication machinery, enhancing processivity and proofreading. The CoV-2 genome is encoded within the +ssRNA and consists of multiple proteins encoded by two overlapping open reading frames (ORFs), ORF1a and ORF1b[3;4] (see Fig. 1.1). ORF1a encodes for nsp1-11, which serve a variety of functions such as RNA capping (nsp10) and potential protein priming (nsp9). Notably, ORF1a also encodes nsp7 and nsp8, which are indispensable cofactors of the RNA-dependent RNA

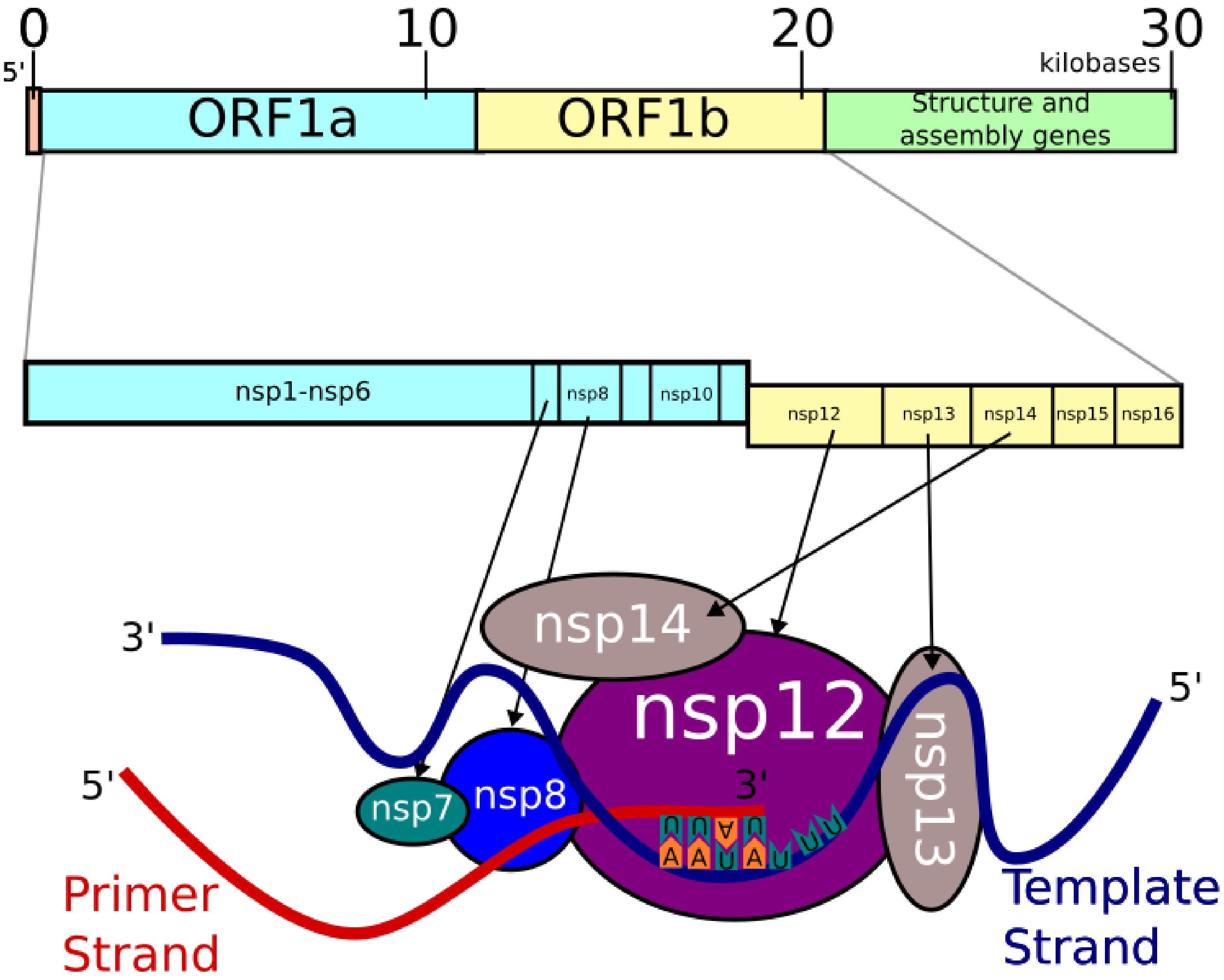


Figure 1.1: The SARS-CoV-2 genome from the ∼30,000 kilobases in the +ssRNA. The ORF1a and ORF1b encode for the nonstructural proteins 1-16. Of key importance are those involved in replication and transcription shown at the *Bottom*. The nsp 12 (RNA dependent RNA polymerase) with cofactors nsp7 and nsp8 make up the core with an RNA duplex: template (blue) amd primer (red). Also shown are the Nsp13 helicase enzyme and Nsp14 exonuclease enzyme.

polymerase (RdRp), nsp12. Experimental studies have demonstrated that nsp12 requires these cofactors for effective RNA polymerization. On the other hand, ORF1b encodes for nsp12-16, including the crucial RdRp (nsp12), the nsp13 helicase, and nsp14 exonuclease, the latter of which proofreads the RNA and can excise improper base pairs. Following the ORFs, the 3' end of the genome encodes structural proteins that encapsulate the +ssRNA and facilitate self-assembly into viral particles. While only a subset of nsps is directly involved in replication and transcription, others are the subject of ongoing investigations to elucidate their functions. In the subsequent section, I will delve into the specifics of CoV-2's RdRp and draw comparisons with polymerases from other viruses.

1.2 The SARS-Cov-2 RNA Dependent RNA Polymerase (nsp12)

The CoV-2 RdRp (or nsp12) is the core engine of the viral replication machinery. Its functional form is achieved through a complex comprising nsp12 itself, the viral cofactors nsp7, and two copies of nsp8, which collectively facilitate RNA template binding and enhance the processivity of RdRp.[5;6] Cryo-EM studies have provided insights into the structural organization of nsp12.[7–11] Across multiple structures solved, it is consistently observed that nsp12 consists of two domains: the N-terminal domain and the polymerase domain (Pol Domain). The N-terminal domain consists of an N-terminal beta hairpin (residues 31 to 50), an extended domain (residues 115-250), adopting a nidovirus RdRp-associated nucleotidyl-transferase (NiRAN) architecture,[12] and an interface region (residues 251-365). The Pol Domain (residues 398-932) adopts a right hand structure with three subdomains: fingers subdomain (residues 366-581 & 621-679), palm (residues 582-620 & 680-815), and thumb (residues 816-932),[13] similar to other RNA Polymerase and RdRps[14–16] (see Fig. 1.2).

1.2.1 The RdRp active site

The CoV-2 RdRp, encompassed within the Pol subdomains, comprises seven conserved structural motifs A-G that are ubiquitous in viral RdRps[17;18] and facilitate nucleotide bind-

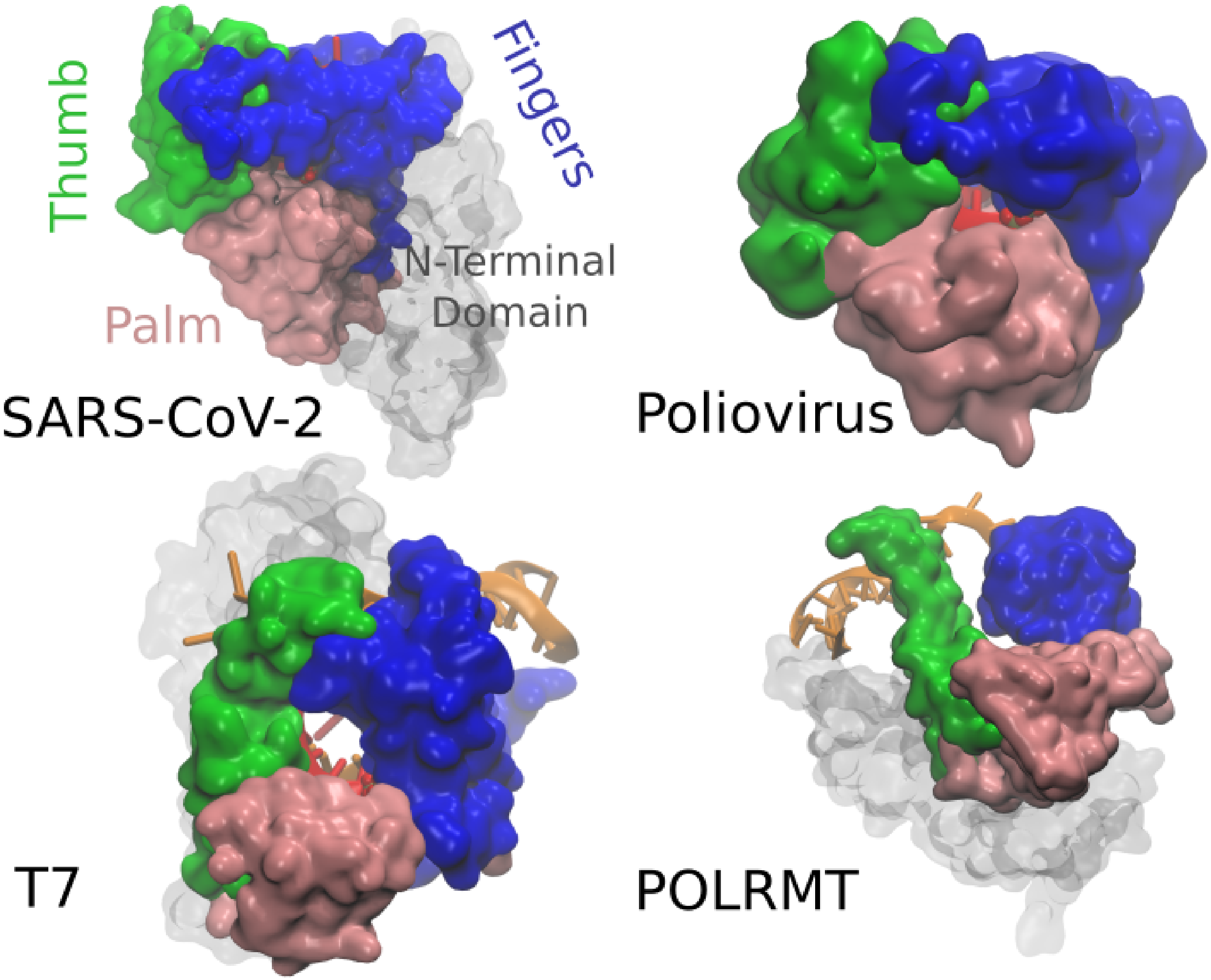


Figure 1.2: The right hand structure of the pol domain from different viruses and the human mitochondria. With the subdomains colored: fingers (blue). palm (pink), and thumb (green). *Upper Left:* The SARS-Cov-2 RdRp pol subdomains and N-terminal domain in transparent gray (PDBid:7BV2). *Upper Right:* The Poliovirus RdRp which structurally similar to CoV-2 (PDBid:5f8j). *Lower Right:* The bacteriophage T7 viral DNA-directed RNA polymerase (PDBid:1s76). *Lower Left:* The DNA-directed RNA polymerase from the human mitochondria (model create from PDBid:4BOC).

ing and catalysis (Fig. 1.3). Motifs A and C contain the conserved aspartic acid residues that are involved in catalysis. In particular, motif C contains the amino acid sequence "SDD" (S759/D760/D761 in CoV-2) which coordinates with two Mg^{2+} ions both needed for catalysis.[19] Motif A contains a conserved D623 that interacts with the 3'-OH group of the ribose sugar subsequent to NTP binding. In addition, motif B is also interacts with the ribose sugar and other sections of the NTP. Motif D was originally believed to provide structural stability, however recent studies on Poliovirius RdRp have shown it could play a role in providing a proper chemical environment for catalysis. The functions of motifs E and G remain unknown, but given motif G's proximity to the template strand, it is likely that it interacts with it. Conversely, motif F engages with the triphosphate moiety of the NTP, and is conjectured to be involved in the initial stages of NTP binding. Finally, motif F interacts with triphosphate moiety of the NTP and is likely involved in the earliest step of NTP binding. Studies in HCV RdRp have shown that the NTP enter via the phosphate first until it reaches the active site, at which point the NTP binds.[20] It is clear that motifs A-G play are crucial for the elongation cycle of the CoV-2 RdRp.

1.2.2 Elongation Cycle

In template-based polymerization or elongation conducted by the CoV-2 RdRp to synthesize RNA chains of thousands of nucleotides, selection of the right (or cognate) nucleotides by the enzyme is highly essential to maintain elongation fidelity over that at equilibrium with Watson-Crick base pairing.[21;22] To do that, an incoming nucleotide binds to the RdRp into the active site (Fig. 1.3), subjected to nucleotide selection at multiple stages, and is then added to the $3'$-end of a synthesizing chain via phosphoryl-transfer reaction. The nucleotide addition cycle (NAC) proceeds in multiple kinetic steps to allow stepwise nucleotide selectivity[22;23] (Fig. 1.4).

The chemical free energy supports the NAC ensure that the polymerase elongates at a non-equilibrium steady state to achieve sufficient speed and accuracy. While the principle

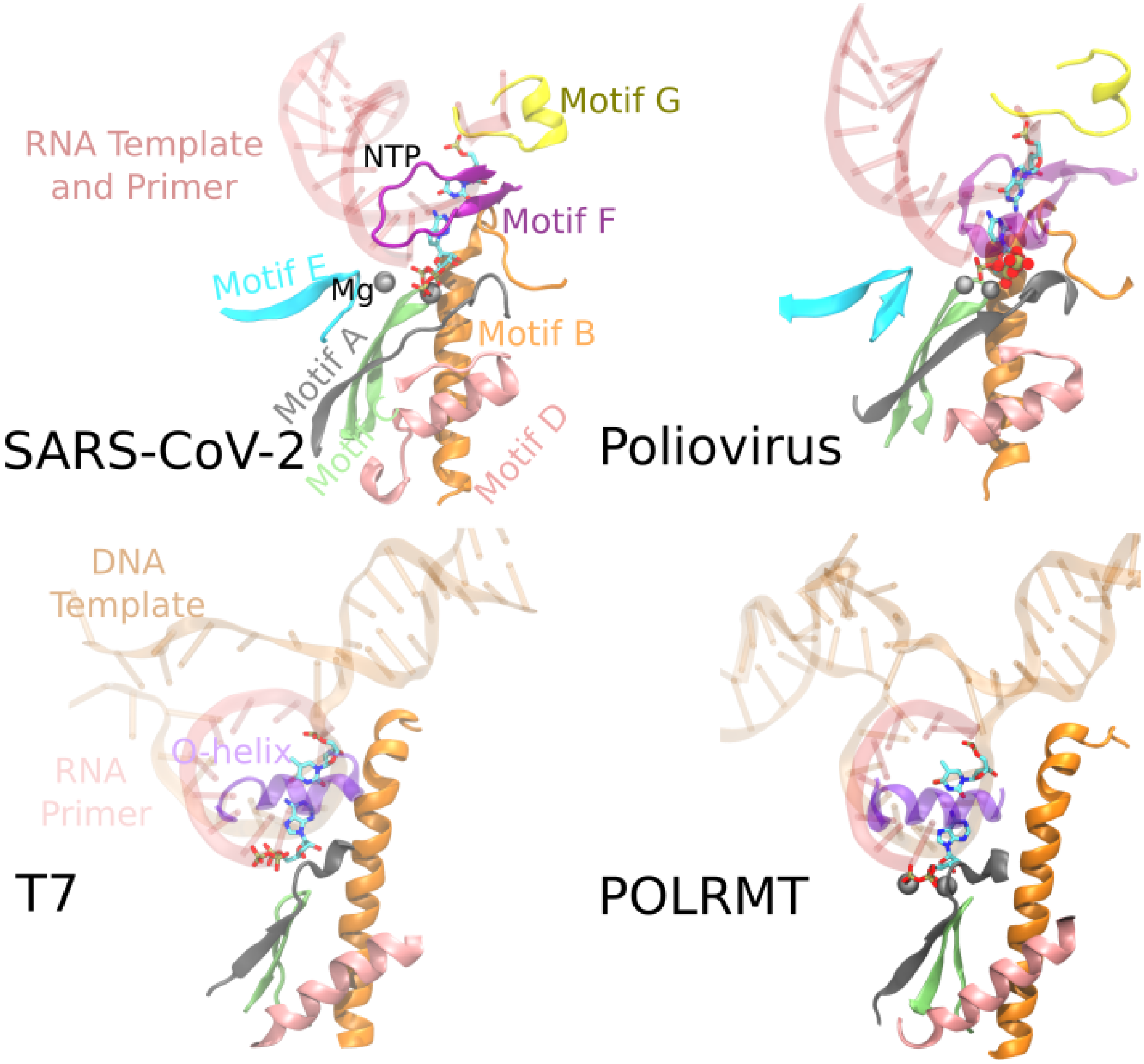


Figure 1.3: The conserved motifs within the Pol domains of various polymerases are essential for their function. As depicted in the *upper panel*, Motifs A-G are structurally conserved across all viral RNA-dependent RNA polymerases (RdRps), and are associated with an RNA duplex (transparent red), magnesium ions (silver), and NTP and template strands (colored by atom name). In contrast, as shown in the *lower panel*, only Motifs A-D are present in DNA-directed RNA polymerases, which are associated with an RNA single strand (red), DNA duplex (transparent orange), magnesium ions (silver), and NTP and template strands (colored by atom name). Motif A (gray) and Motif C (green) contain conserved aspartic acid residues that play a critical role in catalysis. Motif B (orange) consists of residues that interact with the ribose sugar of the RNA. Motif D (pink) is speculated to contribute structural stability or facilitate the appropriate chemical environment for catalysis. Motif G (yellow) is involved in interactions with the template strand. The function of Motif E (light blue) remains elusive. Motif F (purple) is known to interact with the triphosphate moiety of the NTP and is hypothesized to be involved in fidelity control during replication. A structural element in RNA polymerase with a function analogous to Motif F is the O-helix, depicted in transparent purple. Structures are the same as used in Fig. 1.2.

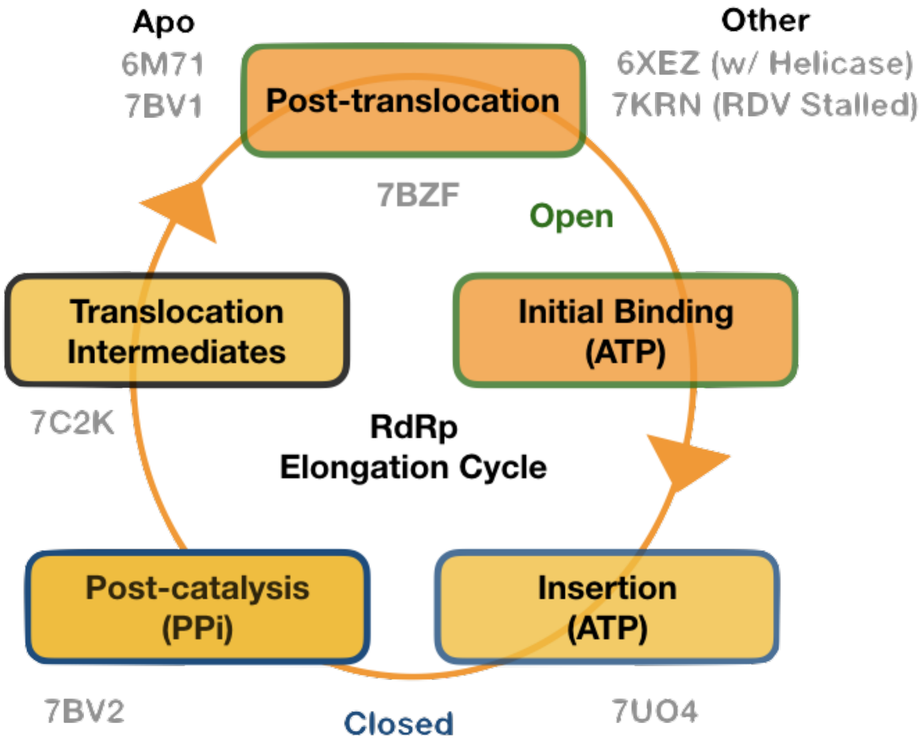


Figure 1.4: This is a schematic representation of the elongation cycle for single nucleotide addition by the SARS-CoV-2 RdRp. Initial binding states or 'open' states are delineated by a green border, while the 'closed' or insertion states are marked in blue. Corresponding cryo-EM structures, identified by their respective PDB IDs, are depicted beneath their respective states and are shown in gray.

of operation seems straightforward, the underlying structural function mechanisms of various polymerase systems necessitate detailed studies for a comprehensive understanding. In particular, we need to focus on the following points: i) the process of screening an incoming nucleotide upon its initial binding to the active site or a nearby region, ii) the mechanism of selection as the nucleotide is progressively inserted into the active site for chemical reaction readiness, and iii) the process of nucleotide discrimination during the chemical reaction, which is typically catalyzed by two metal ions (such as Mg 2+ or Mn 2+ ions).[19] In addition, the integrated nucleotide can introduce additional selectivity after the chemical addition.

This dissertation primarily focuses on points i) and ii). Once the nucleotide is in the active site we refer to it as initial binding or "open" state. A subsequent conformational change leads to be an insertion or "closed" state. This conformational change has been shown to be a subtle ∼2Å shift of motifs in the palm subdomain becoming closer to the fingers subdomain in Poliovirus[16] and Enterovirus.[24] It is hypothesized that this pre-chemistry step is an essential fidelity checkpoint[22;25] and possible rate determining step. More specifically, this dissertation focuses on the insertion of four NTPs: ATP, Remdesivir-TP (RDV-TP a nucleotide analogue to adenine), dATP, and GTP.

Chapters 2 and 3 delve into ATP and RDV-TP. RDV-TP was the first FDA approved therapeutic to treat COVID infection before vaccines were developed. Chapters 4 and 5 shift the focus to non-cognate nucleotides: dATP, which has a sugar mismatch due to a missing 2'-OH group, and GTP, which presents a base mismatch to the model containing a uracil template.

The understanding of how a proper nucleotide is selected is grounded in atomistic scale detail. Moreover, the transition from an 'open' to 'closed' state is projected to occur on the millisecond timescale. Therefore, to unravel these intricate details, both equilibrium and non-equilibrium molecular dynamics simulation methods were employed. The specifics of these methods will be discussed in the following section.

1.3 Molecular Dynamics Simulations

The research described in this thesis utilized various computational tools, primarily Molecular Dynamics (MD) simulations. MD simulations involve a classical approximation of a molecular system which gives us the dynamics at an atomistic scale. The initial configuration for MD simulations can typically be obtained from x-ray crystal or cryo-EM structures. Using these coordinates one can solve the equations of motion on an atom $\overrightarrow{r}_i$ by determining the acceleration from the forces on an atom $\overrightarrow{F_i}$ through newtons second law:

$$\overrightarrow{F_i} = m_i a_i = m_i \frac{\partial^2 \overrightarrow{r}_i}{\partial^2 t} \tag{1.1}$$

Here, m_a represents the mass of an atom, and t represents time. The forces acting on atoms $(\overrightarrow{F_i})$ can be determined by calculating the negative gradient of the potential energy function:

$$\overrightarrow{F_i} = -\nabla V(r_{ij}) \tag{1.2}$$

The potential energy function for a biomolecular system incorporates essential chemistry concepts such as covalent bond, dihedral angles, van der Waals interactions, and electrostatics.

$$\begin{aligned} V = \sum_{bonds} k_r(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi[1 + cos(n\phi + \phi_0)] \\ + \sum_{atom\ i} \sum_{j \neq i} 4\epsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{i,j}}{r_{ij}} \right)^{6} \right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon_0 r_{i,j}} \end{aligned} \tag{1.3}$$

The gradient of this potential energy surface is commonly known as a force field. The AMBER force field, as well as the CHARMM force field, are popular choices, each involving different parameters (e.g., K_r, r_0, K_θ, θ_0) that are determined by fitting experimental data and quantum mechanics calculations. While classical force fields have limitations, particularly when dealing with highly charged or small molecules, these limitations are well recognized within the scientific community.[26;27] Extensive efforts are being made to develop efficient polarisable force fields[28;29] and improved small molecule force fields.[30] Despite these limitations, classical force fields allow efficient and reasonably accurate simulations of biological molecules in solution. The present dissertation employs the AMBER force field parameters.

1.3.1 Brief History of AMBER force field updates

In this dissertation, a portion is dedicated to the development and analysis of a nucleotide analogue force field (ff), which will be elaborated upon in a subsequent chapter. To effectively achieve this and foster a comprehensive understanding, I surveyed the literature, which has numerous AMBER force fields, often having confusing nomenclatures. In the interest of future scholars, I have composed a concise and cogent overview.

The first big step in making macromolecular ff was in 1983, when the CHARMM program came out with its own ff.[31] In 1995, the AMBER package made another big step by adding parameters that work well with nucleic acids, and this ff was named AMBER 95.[32] While there were small updates to AMBER 95 the next popular ff was an update to AMBER99 titled AMBER99SB[33] which only updated protein parameters. This was followed up with AMBER99SB-ILDN[34] which specifically improved side-chain torsion potentials of two residues. Currently, the newest and most commonly used protein ff is AMBER14sb,[35] which was used in the research described in later chapters. Although there exists a more modern release, AMBER ff19sb,[36] this ff requires the use of the more accurate OPC[37] water model. The use of OPC increases the simulation cost by $\sim$33% as opposed to the more typically used tip3p.[38] This increase in cost has led to this ff not being widely adopted.

The first notable enhancements to nucleic acid parameters were introduced with the release of the parmbsc0 force field,[39] which was later followed by the parmbsc1 ff.[40] Currently, parmbsc1 is the most up-to-date set of nucleic acid parameters. The primary focus of these updates has been on refining the torsional values. Concurrently with the development of bsc1, an alternative set of nucleic acid parameters named OL15 was developed.[41] However, comprehensive comparisons between parmbsc1 and OL15[27] have revealed that both ff yield similar results.

Until now, the ff discussed have primarily been geared towards biomacromolecular structures. However, small molecules necessitate a different approach. To cater to a broader spectrum of small molecules, Amber introduced the General Amber Force Field (GAFF),[42] encompassing a wide array of common functional groups and atom types. Despite this, there are still numerous functional groups not covered by GAFF that may be of interest, indicating a need for further research in this area.

1.3.2 Biased MD: Umbrella Sampling

One of the most significant improvements in the computational resources available for simulations is the capability to offload calculations onto Graphics Processing Units (GPUs).[43] The incorporation of GPUs has significantly extended the feasible lengths of equilibrium MD simulations. At the start of my dissertation work, the limiting factor was the microsecond timescale. Now, we are slowly reaching into the sub-millisecond timescale.[44] However, many problems of interest, such as the open-to-close conformational change described in section 1.2.2, fall within the millisecond to second timescale and thus remain largely unreachable. Even if we were to run a 1-2 ms simulation and observe a conformational change, we would be left with only a single data point, introducing a sampling issue. To overcome these limitations, non-equilibrium methods have been developed to extend our sampling capacity beyond these timescales. These methods include, but are not limited to, Umbrella Sampling,[45] Metadynamics,[46] and Gaussian Accelerated MD.[47]

The effectiveness of most biasing or non-equilibrium methods hinges on the correct selection of a reaction coordinate (RC) or collective variable (CV). The CV, which ideally captures the phenomenon under study, often poses the greatest challenge in the process.[48] To better illustrate the use of a CV, consider an MD simulation at temperature T samples

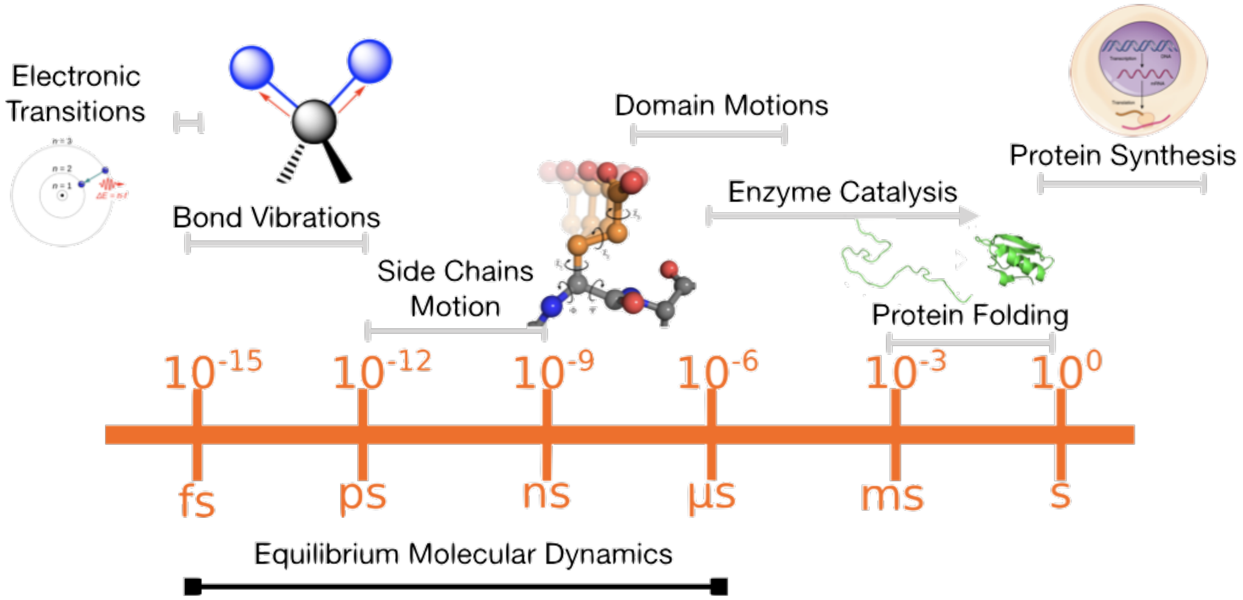


Figure 1.5: The biophysical timescales with relevant motions for reference. Equilibrium molecular dynamics simulations (black line) are limited to the microsecond to sub-millisecond timescale.

conformations from canonical ensemble:

$$P(q) \propto e^{-\frac{V(q)}{k_B T}} \tag{1.4}$$

In this case, the Boltzmann-weighted potential energy governs the probability distribution. Thus, the free energy or probability distribution of a specific CV, denoted as s, would be described by:

$$\begin{aligned} P(s) &\propto \int dq\ e^{-\frac{V(q)}{k_B T}} \delta(s - s(q)) \\ F(s) &= -k_B T \log P(s) \end{aligned} \tag{1.5}$$

As previously mentioned, sampling the space described by CV s can be challenging, even with a perfect CV. Hence, we can opt to sample a biased probability distribution or free energy space based on the potential of the CV. By adding a term $V(s(q))$ that only affects the CV, we have:

$$\begin{aligned} P'(s) &\propto \int dq\ e^{-\frac{U(q)+V(s(q))}{k_B T}} \delta(s - s(q)) \\ &\propto e^{-\frac{V(s(q))}{k_B T}} P(s) \\ F'(s) &= -k_B T \log P'(s) \\ &= F(s) + V(s) + C \\ F(s) &= F'(s) - V(s) + C \end{aligned} \tag{1.6}$$

To recalculate the unbiased information (free energy), we can re-weight w a biased trajectory(s):

$$\begin{aligned} P(q) &\propto P'(s) e^{-\frac{V(s(q))}{k_B T}} \\ w &\propto e^{-\frac{V(s(q))}{k_B T}} \end{aligned} \tag{1.7}$$

But even calculating the re-weighting factor is a nontrivial task. However, if we have a good idea on the description of the CV and where the barrier lies a path can be generated which you can then apply a harmonic restraint on in the form:

$$\begin{aligned} V(s) &= \frac{k}{2}(s - s_0)^2 \\ P(q) &\propto P'(s) e^{-\frac{k(s(q)-s_0)^2}{2k_B T}} \end{aligned} \tag{1.8}$$

Here in lies the beauty of the U.S. method. Given that it uses a harmonic restraint along a CV, this leads to a set of simple algorithms applicable for re-weighting such as WHAM. For the purpose of this thesis, I have used the Umbrella Sampling method.[49]

Chapter 2

ATP and Remdesivir: A Rigorous Examination of Initial Binding and Insertion States

This chapter centers on the initial binding and insertion states of ATP and drug analogue RDV-TP, which are the bases for all of the simulations . At the time of this study,[50] research on SARS-CoV-2 RdRp was in its infancy. Only two cryo-EM structures were available, and they did not represent our states of interest. Furthermore, given that RDV is an adenine analogue and nonstandard residue, no existing force field was applicable. We, therefore, constructed our own. Meticulous attention was paid to validating our constructed models and our equilibrium results. With the knowledge gained from this process, we calculated the free energy of insertion using more expensive non-equilibrium methods. Hence, this chapter is fundamental to the comprehension of the remaining work.

2.1 Introduction

Due to its critical role in the viral RNA synthesis and highly conserved core structure, the viral RdRp serves a highly promising antiviral drug target for both nucleotide analogue and non-nucleoside inhibitors.[51] Remdesivir (or RDV), the only US-FDA proved drug (named VEKLURY) so far treating COVID-19,[52] works as a prodrug that is metabolized into a nucleotide analogue to compete with natural nucleotide substrates of RdRp to be incorporated into viral RNA gnome to further terminate the RNA synthesis.[53;54] As a broad-spectrum anti-viral compound, RDV was developed originally for treatments of Ebola virus disease (EVD),[55] and then applied for infections of middle east and severe accurate respiratory syndrome coronavirus (MERS-CoV and SARS-CoV),[56] which are both close relatives to the currently emerged novel coronavirus (SARS-CoV-2) causing COVID-19. Recent in-vitro and in-vivo studies on RDV impacts to the viral RdRp function have confirmed the RDV analogue incorporation and inhibition during the viral RdRp replication, in particular, in SARS-CoV-2.[57–60] The existing evidences have consistently suggested that the active triphosphate form of RDV (RDV-TP) binds competitively with the natural substrate, i.e., adenosine triphosphate (or ATP), to the viral RdRp and the incorporation leads to a delayed chain termination.[58] Such an analogue incorporation and consequent chain termination indicate that RDV-TP can successfully evade from both nucleotide selectivity of the viral RdRp as well as the proofreading function from ExoN in coordination with RdRp in the coronavirus replication.[53;61]

Nucleotide selectivity of the RdRp or polymerases in general serves as a primary fidelity control method in corresponding gene transcription or replication, i.e., during the template-based polymerase elongation.[21;25;62] The selectivity indeed proceeds throughout a full nucleotide addition cycle (NAC), consisting of nucleotide substrate initial binding, insertion to the active site, catalysis, product (or pyrophosphate) release, and together with the polymerase translocation.[63] To be successfully incorporated, the antiviral nucleotide analogue needs to pass almost every fidelity checkpoint in the polymerase NAC.[22;23] In coronaviruses with large genome sizes, proofreading conducted by an exonuclease (or ExoN) protein further improves the RNA synthesis fidelity.[61] Correspondingly, the nucleotide analogue drug need further evade from the ExoN proofreading to ultimately terminate the RdRp elongation. Although RDV succeeds as a nucleotide analogue drug to interfere with the CoV-2 RdRp function, as being demonstrated in vivo and in vitro, the underlying structural dynamics mechanisms on how that being achieved are still to be determined, and *in silio* approaches may particularly help. Recent modeling and computational efforts have been made to approach the underlying mechanisms of the RDV-TP binding and incorporation to the CoV-2 RdRp, from molecular docking[64] and binding free energy calculation upon the nucleotide initial association,[65] to nucleotide addition together with potential ExoN proofreading activities.[66] Nevertheless, those studied structural systems were still made by constructing homology model of the SARS-CoV-2 RdRp according to a previously resolved structure of the SARS-CoV RdRp.[67] Upon very recent high-resolution cryo-EM structures being resolved on the SARS-CoV2 RdRp (the non-structural protein or nsp12), with and without incorporation of RDV,[7;11] it becomes highly desirable to conduct all-atom modeling and molecular dynamics (MD) simulations directly on the CoV-2 RdRp structure, so that to probe how RDV succeeds at binding and inserting into the RdRp active site, despite of existing nucleotide selectivity of RdRp to be against non-cognate nucleotide species.[62]

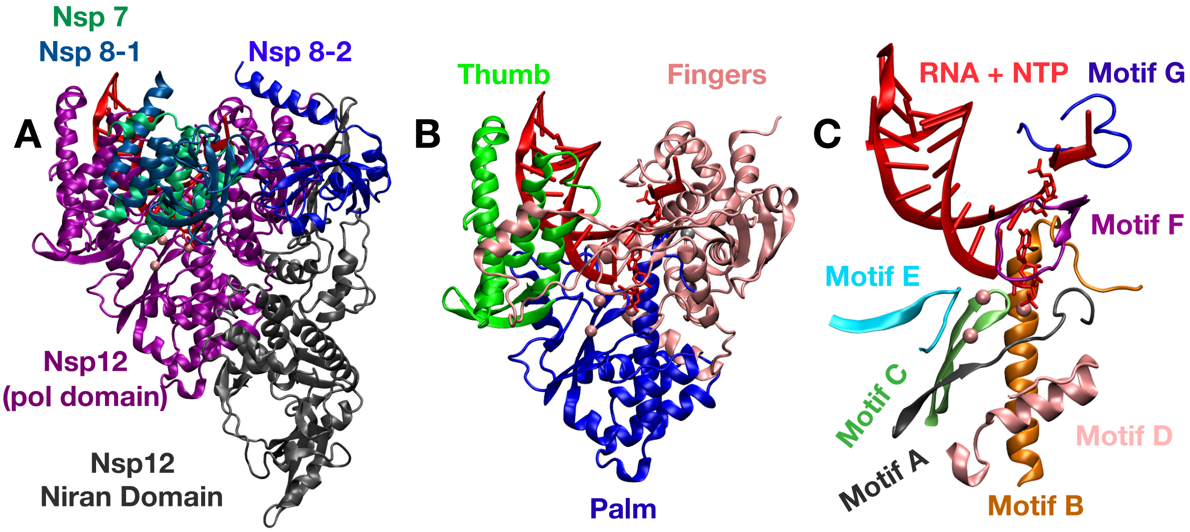


Figure 2.1: SARS-CoV-2 RdRp elongation complex with an incorporated remdesivir (RDV) in the closed state (based on PDB:7BV2[11]). **A** The two main domains (N-terminus domain in grey, polymerase or pol domain in purple) of the RdRp along with the three cofactors (nsp8's in blue and nsp7 in green). **B** The pol domain consists three subdomains, the thumb (green), fingers (pink), and palm (blue). RNA (red) is shown along with incoming NTP and +1 template nt (red licorice). **C** Motifs A-G within the pol domain.

The high-resolution structures of SARS-CoV-2 RdRp or nsp12 were obtained in complex with accessory protein nsp7 and nsp8, which are supposed to assist processivity of the replication/transcription machinery along the viral RNA[68] (see Fig. 2.1A). The core RdRp (residue 367-920, excluding the N-terminal NiRAN and interfacial region) adopts a handlike structure, consisting of fingers, palm, and thumb subdomains, similar to other single-subunit viral RNA polymerases (RNAPs) and family-A DNA polymerases (DNAPs).[69–71] There are seven highly conserved structural motifs shared by RdRps, located in the palm (A-E) and fingers (F-G) subdomains (Fig. 2.1B). In general, when there is no substrate bound, the

RdRp active site adopts an open conformation. A nucleotide substrate can bind to the active site in the open conformation, and inserts into the active site to reach a closed conformation, as the nucleotide is stabilized or to be ready for the catalytic reaction.[13;16] In the recently resolved SARS-CoV-2 RdRp structures, both the open and closed conformation state of the active site were captured, with the former in the absence of the substrate,[7] and the latter captured with an RDV analogue already incorporated to the end of the synthesizing RNA chain (i.e., in post-catalytic or product state).[11] In order to probe how a nucleotide or analogue binds and inserts to the RdRp active site, we accordingly constructed both an open (i.e. substrate initial binding) and a closed (substrate insertion) structural complex of the CoV-2 RdRp, based on the newly resolved structures (PDB: 7BTF[7] and 7BV2[11]) (see Fig. 2.1C for a closed form).

2.2 Computational Details

2.2.1 Constructing Initial Binding & Insertion structures

High-resolution Cryo-EM structures for CoV-2-RdRp's elongation complex are available in a post-catalysis state with the RDV analogue incorporated (PDB:7BV2).[11] Since this is a tertiary elongation complex of RdRp (with a full length nsp12) captured together with RNA strands (template and primer) and the RDV analog incorporated (post-catalysis or product state; PDB: 7BV2), presumably in the active-site closed state, we built an RDV-TP insertion model (pre-catalytic) of the CoV-2 RdRp directly using this tertiary complex, only replacing the incorporated RDV analog at the 3'-end of the RNA primer strand by a pre-catalytic RDV-TP. The three Mg^{2+} ions present in the product state are taken as initial positions for our model built for. This same model was then used to construct ATP by replacing RDV-TP with an aligned ATP.

The active-site open structure of the CoV-2 RdRp for NTP initial binding was obtained from the first determined cryo-EM structure[7] (PDB: 7BTF). Then RDV-TP was placed to the active site of the open state structure (along with the RNA template and primer strand) by aligning the RdRp structure from the tertiary RDV-TP insertion complex (the closed one constructed above) with that of the open one, and then shifting the RDV-TP and RNA strands from the tertiary complex to the open state structure accordingly. Additionally, the Mg^{2+} ions in the initial binding state were kept similarly as in the insertion state. Followed, the modeled structural complex would be subject to MD simulation equilibration. Additionally, two nsp8 N-terminals were cleaved and shorted by 11 residues to avoid instabilities.

Missing residues were added using MODELLER 9.24[72] and an apo nsp12 structure as a reference (PDB:7BTF).[7] PDBID:7btf is missing the following residues in the N-terminus domain: 1-30, 51-68, 75, 103-111, 895-906, and the following in the thumb sub-domain 920-932. Missing residues were completed using MODELLER 9.24[72] with the apo structure PDBID:6M71 which is only missing residues 1-4.

Histidine protonation states were predicted using PDB2PQR[73] and PROPKA3[74] followed by visual inspection:

- HID: 75 99 113 133 256 347 355 362 439 572 599 613 810 816 882 898
- HIE: 82 295 309 381 642 650 725 752 872 892
- Residue 295 and 642 are manually selected due to their orientation with Zn ion such that the proton is not oriented near the metal.

2.2.2 Generating a Force Field for RDV-TP

A force field was generated for RDV, with partial charges calculated by following the formalism used in AMBER nucleic acid forcefields.[32] RDV-TP 3' and 5' terminals were truncated, and replaced with terminal hydroxyl groups (Fig. 2.2).

A Hartree-Fock calculation at the level of HF/6-31G* was set to perform geometric optimization and a self-consistent calculation to obtain an electro-static potential for constrained charge fitting. Using the two-stage Restrained Electrostatic Potential method,[75] partial atomic charges for the RDV were generated. During Restrained Electrostatic Poten-

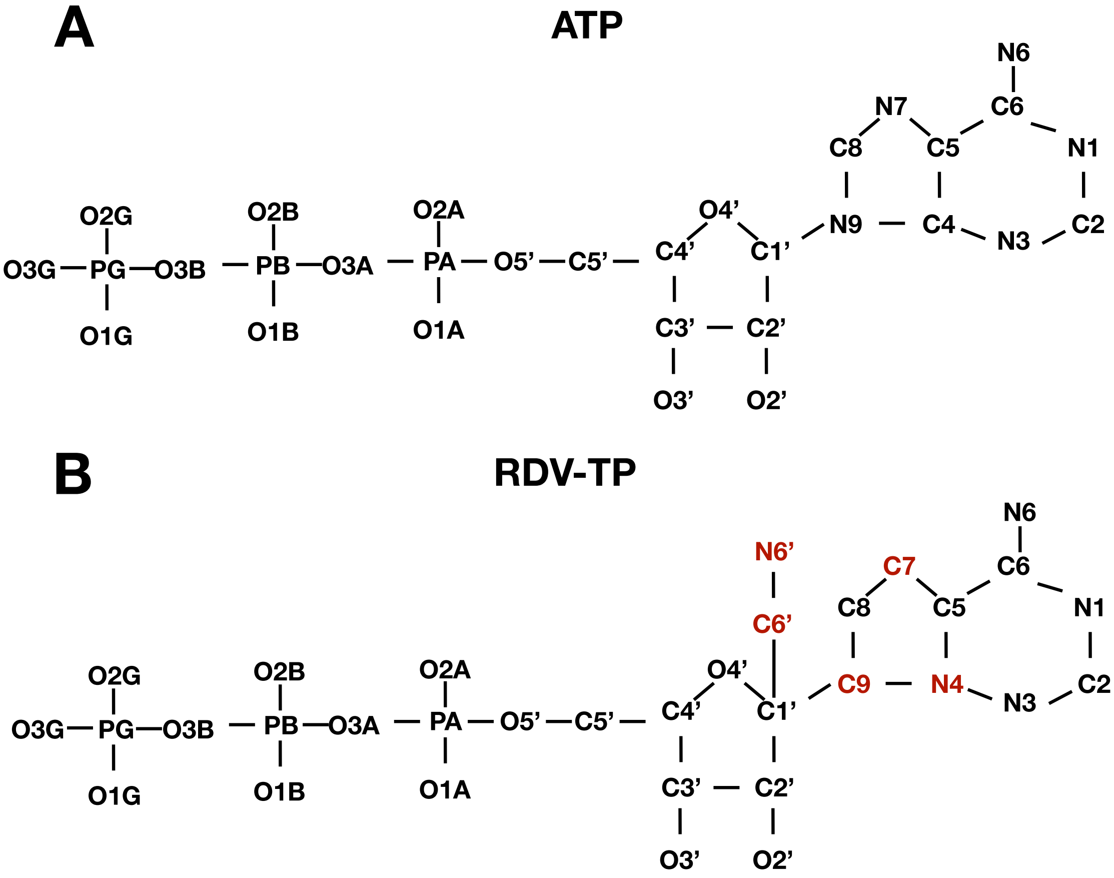


Figure 2.2: Comparing **A.** Adenosine Triphosphate (ATP) and **B.** Remdesivir Triphosphate (RDV-TP) heavy atom molecular structures. Hydrogens are omitted for a clear representation. Atoms colored in red highlight the differences in RDV from ATP. For partial charge calculation the RDV-TP is truncated at the O5' with the addition of a hydrogen H5T. H3T is the hydrogen atom bonded to the O3' oxygen (see Table. 2.1).

tial method (RESP[75]) the 3' 5' hydroxyl atomic charges are constrained[32] to O5'= -0.6223e, H5T=0.4295e, O3'=-0.6541e, H3T=0.4376e, partial atomic charges for the truncated Remdesivir are generated (see Table. 2.1).

Table 2.1: Summary of Partial Charges used for the RDV-TP force field compared with ATP. Charges are separated by section of the NTP.

Atom (PolyP)	ATP	RTP	Atom (Sugar)	ATP	RTP	Atom (Base)	ATP	RTP
O5'	-0.59870	-0.59870	C5'	0.05580	0.039981	N9	-0.02510	N.A.
PA	1.25320	1.25320	H5'1	0.06790	0.085276	C9	N.A.	-0.118619
O1A	-0.87990	-0.87990	H5'2	0.06790	0.085276	C8	0.20060	-0.228326
O2A	-0.87990	-0.87990	C4'	0.10650	0.083427	H8	0.15530	0.199975
O3A	-0.56890	-0.56890	H4'	0.11740	0.065203	N7	-0.60730	N.A.
PB	1.38520	1.38520	O4'	-0.35480	-0.332867	C7	N.A.	-0.259805
O1B	-0.88940	-0.88940	C1'	0.03940	0.130365	C5	0.05150	-0.394730
O2B	-0.88940	-0.88940	H1'	0.20070	N.A.	C6	0.70090	1.014028
O3B	-0.53220	-0.53220	C6'	N.A.	0.461023	N6	-0.90190	-1.042464
PG	1.26500	1.26500	N6'	N.A.	-0.505959	H61	0.41150	0.443695
O1G	-0.95260	-0.95260	C3'	0.20220	0.329872	H62	0.41150	0.443695
O2G	-0.95260	-0.95260	H3'	0.06150	0.076195	N1	-0.76150	-0.863108
O3G	-0.95260	-0.95260	C2'	0.06700	-0.074121	C2	0.58750	0.630021
			H2'1	0.09720	0.146103	H2	0.04730	0.076123
			O2'	-0.61390	-0.626760	N3	-0.69970	-0.744123
			HO'2	0.41860	0.462358	C4	0.30530	N.A.
			O3'	-0.65410	-0.65410	N4	N.A.	0.603011
			H3T	0.43760	0.437600			

Torsional parameters were taken from Parmbsc1 when applicable and the general AMBER force field (GAFF).[76] In order to select appropriate torsional from the AMBER force field library, atom types need to be selected. In general the majority of atom types were kept the same as the adenosine where possible from the AMBER force field (see Fig. 2.2).[40] The swapped and additional atoms (nitrile functional group) used the following atom types for RDV: C9:CK, C7:CK, N4:na, C6':c1, and N6':n1. Where the lower case atom types are taken from the Generalized AMBER Force Field (GAFF).[76] The RDV force field parameters were put together using antechAMBER.[77] All input files and scripts used to generate the RDV force field are in the **Appendix**.

2.2.3 Docking

In order to test whether the above constructed initial binding complex was reasonable, we also performed docking of RDV-TP (or RTP below) and ATP as ligands onto the open structural complex of RdRp (nsp12+nsp7+ns8 and RNA together as the receptor), using AutoDock Vina software (see Fig. 2.4).[78] The receptor complex was prepared by deletion of water molecules, addition of hydrogen molecules and by computing Kollman charges. The ligands (RTP and ATP) were prepared by computing Gasteiger charges. A grid box (x=40 Å, y=40 Å, z=40 Å) is specified around the active site for the search space on the receptor within which various positions of the ligand are to be considered. An energy range of 4 and exhaustiveness of 8 were assigned. Conformations with lowest binding free energetic scores are considered most stable or optimal.

2.2.4 Equilibrium Simulation Details

All MD simulations were performed using Gromacs 2019 package[79] with the Amber14sb protein force field[35] and Parmbsc1 nucleic acid parameters.[40] For the NTPs, triphosphate parameters calculated previously were used.[80] Each of the RdRp complexes was solvated with explicit TIP3P water[38] with a minimum distance from the complex to the wall set to 15Å, resulting in an average box size of 15.7nm x 15.7 nm x 15.7 nm (see Fig. 2.3). Sodium and chloride ions were added to neutralize the systems and make the salt concentration 100mM. Three magnesium ions were kept from the cryo-EM structures (though

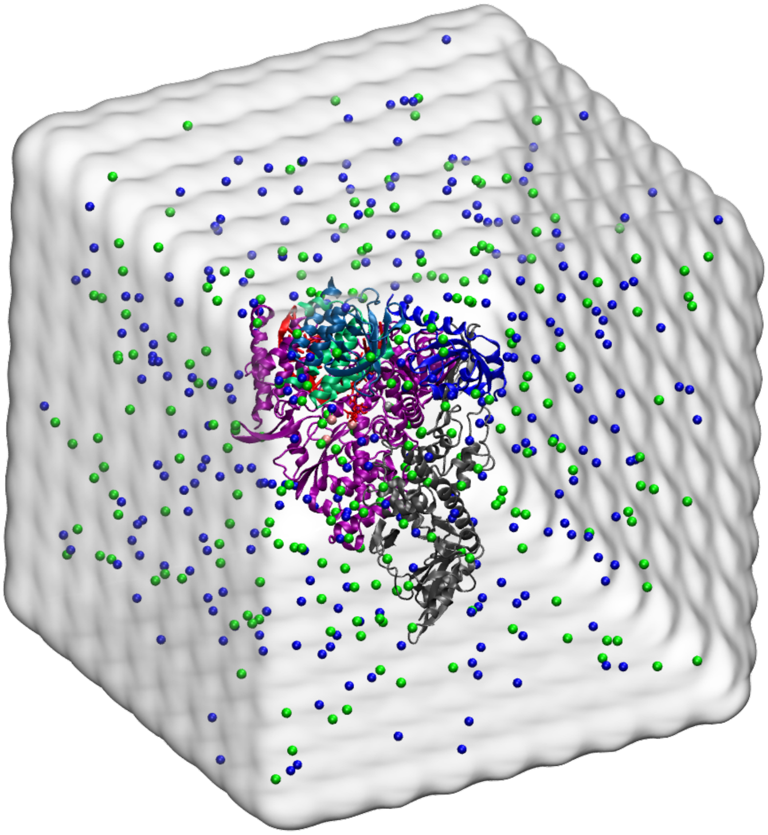


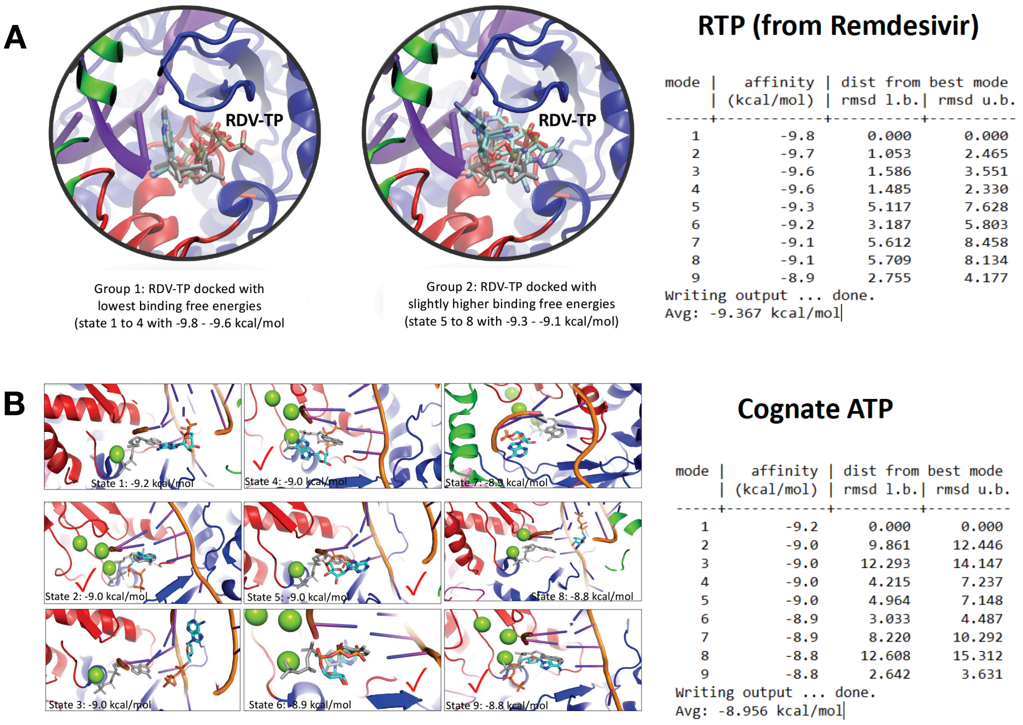
Figure 2.3: All-atom molecular dynamics simulation box. The size of the box on average: 15.7nm x 15.7 nm x 15.7 nm, containing an average of 382,000 atoms.

only two are supposed to be catalytically relevant).[11] The full simulation systems contained on average about 382,000 atoms. For all simulations, the cut-off of van der Waals (vdw) and the short range electrostatic interactions were set to 10Å. Particle-mesh-Ewald (PME) method[81;82] was used to evaluate the long-range electrostatic interactions. Timestep was 2 fs and the neighbor list was updated every 10 steps. Temperature was kept at 310 K using the velocity re-scaling thermostat.[83] Pressure was kept at 1 bar using Berendsen barostat[84] during equilibration. Each initial system was minimized for a maximum of 50000 steps using steepest-descent algorithm, followed by a 2-ns NVT MD simulation with all the heavy atoms in the system fully constrained. Next a 2-ns NPT simulation along with the same constraints was performed. Constraints were released in 1-ns intervals in the following order:

RNA, nsp8/nsp-7, nsp12/NTP/metal ions. In total for each initial binding and insertion states, ten 100ns equilibration trajectories for ATP and RDV-TP systems were launched independently for a total of $4\mu s$ of simulation time.

Equilibrium trajectories were analyzed via visual inspection and measurement of proper NTP and template nt + 1 base pairing using the distance of heavy atoms. Proper equilibration was gauged by measuring the RMSD of each complex segment or subdomain.

2.3 Results



RTP (from Remdesivir)

```
mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1         -9.8      0.000      0.000
   2         -9.7      1.053      2.465
   3         -9.6      1.586      3.551
   4         -9.6      1.485      2.330
   5         -9.3      5.117      7.628
   6         -9.2      3.187      5.803
   7         -9.1      5.612      8.458
   8         -9.1      5.709      8.134
   9         -8.9      2.755      4.177
Writing output ... done.
Avg: -9.367 kcal/mol
```

Cognate ATP

```
mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1         -9.2      0.000      0.000
   2         -9.0      9.861     12.446
   3         -9.0     12.293     14.147
   4         -9.0      4.215      7.237
   5         -9.0      4.964      7.148
   6         -8.9      3.033      4.487
   7         -8.9      8.220     10.292
   8         -8.8     12.608     15.312
   9         -8.8      2.642      3.631
Writing output ... done.
Avg: -8.956 kcal/mol
```

Figure 2.4: Docking of RDV-TP (or RTP) and ATP onto a modeled initial binding (active-site open) structure of SARS-CoV-2 RdRp (PDB: 7BTF). **A.** RDV-TP docking show comparatively stabilized docking structures (grouped into two). The palm, fingers, thumb subdomains are shown in red, blue, and green, and RNA in violet. The modeled RDV-TP (positioning from PDB: 7BV2) is shown in gray, and the docking structures of RDV-TP are shown with colored atoms. **B.** ATP docking shows diverse configurations and less stabilized configurations. The obtained docking energetics are listed on the right side for both systems (using AutoDock)[78]

To validate our models we conducted equilibrium MD simulations on the optimized docking (Fig. 2.4) complex of ATP. The results (Fig. 2.5) show that even upon the docking and equilibration, the stabilized configurations of ATP still converge to be very close to the initial modeled ones. Interestingly, even we chose another reference structure in docking (e.g. using the pre-insertion structure of T7 RNAP[14]), we could still obtain a docking configuration overlapping well with the insertion ATP. Hence, it justifies that the constructed ATP and RTP initial binding or the active-site open RdRp structural complexes are reasonable. A further comparison show that ATP binding configuration in our constructed open form RdRp complex is similar to that being captured in the PV RdRp.

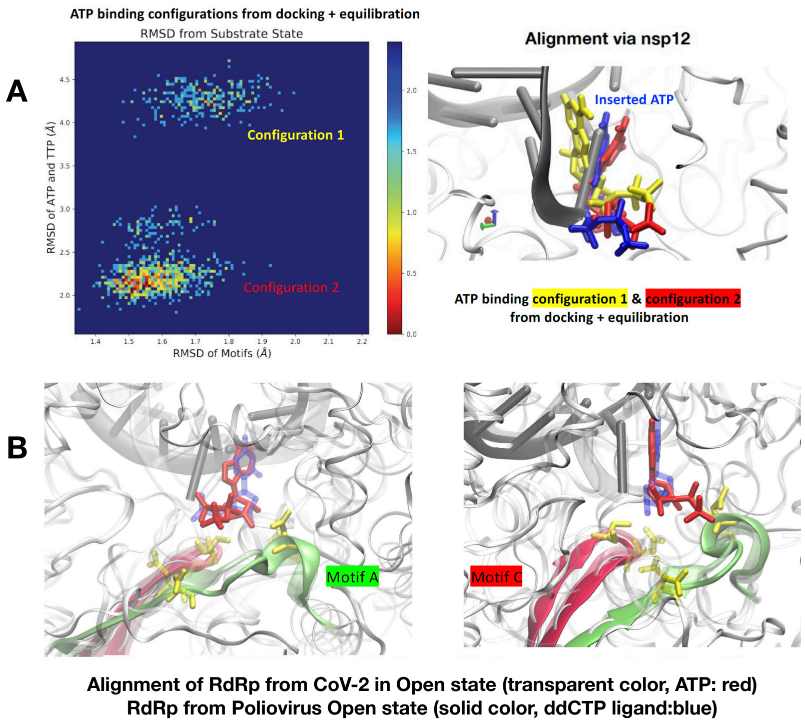


Figure 2.5: Examining ATP binding configurations to the open form active site CoV-2 RdRp. **A.** The MD equilibration of the optimal docking complex of ATP to the CoV-2 RdRp structure. The equilibrated configurations were measured by RMSDs for both structural motifs (A-G) and ATP+template nt (uracil) with respect to the substrate insertion complex. Two dominant configurations of ATP have been identified, both of which are quite close to the insertion configuration (as our modeled open or initial binding complex of ATP to the RdRp, see Methods 2.1). **B.** The alignment of our modeled ATP bound open equilibrated form CoV-2 RdRp with that of the poliovirus (PV) RdRp, shown in two views for better visualization.

Upon validating the active-site closed state complexes for the RDV-TP and ATP insertion, respectively, and then constructing the active-site open state complex to allow the substrate to bind initially (see 2.2.1), we conducted equilibrium all-atom MD simulations for the closed and open complex systems, bound with RDV-TP or ATP (see Fig. 2.6). Base pairing between RDV-TP or ATP with the +1 template nt (Uracil) can well be maintained in the closed state (see Fig. 2.6 and Fig. 2.7). In the open state the base pairing between the RDV-TP or ATP with the template nt appears less or slightly less stabilized. Interestingly, base stacking configuration between RDV-TP and the template nt can be frequently captured, in which the nt base usually stacks upstream relative to RDV-TP (see Fig. 2.6).

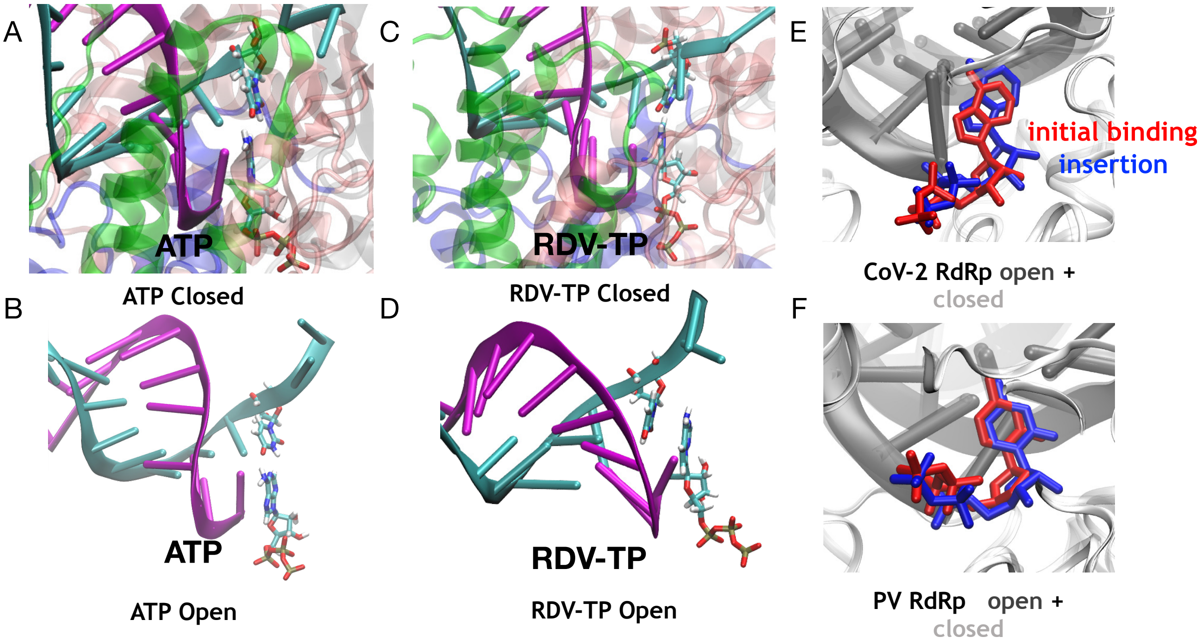


Figure 2.6: Modeled insertion structural complexes of SARS-CoV-2 for RDV-TP and ATP. Left and Center: The active site views with inserted ATP and RDV-TP shown at the end of equilibrium simulations for the insertion (A & C) and initial binding (B & D) states. Right: The open and closed RdRp structures aligned (E), with ATP initial binding and inserted shown, respectively. The CoV-2 RdRp is shown in comparison with previously studied RdRp from Poliovirus (PV) (F) (PDBs: 3ola and 3ol7).[16]

Upon MD equilibration of the initial binding open-state RdRp complex with ATP (∼100 ns; see Fig. 2.8 for RMSD), we found that ATP shows primarily the base pairing initial binding configuration with the +1 template nt. The base pairing interactions seem to be

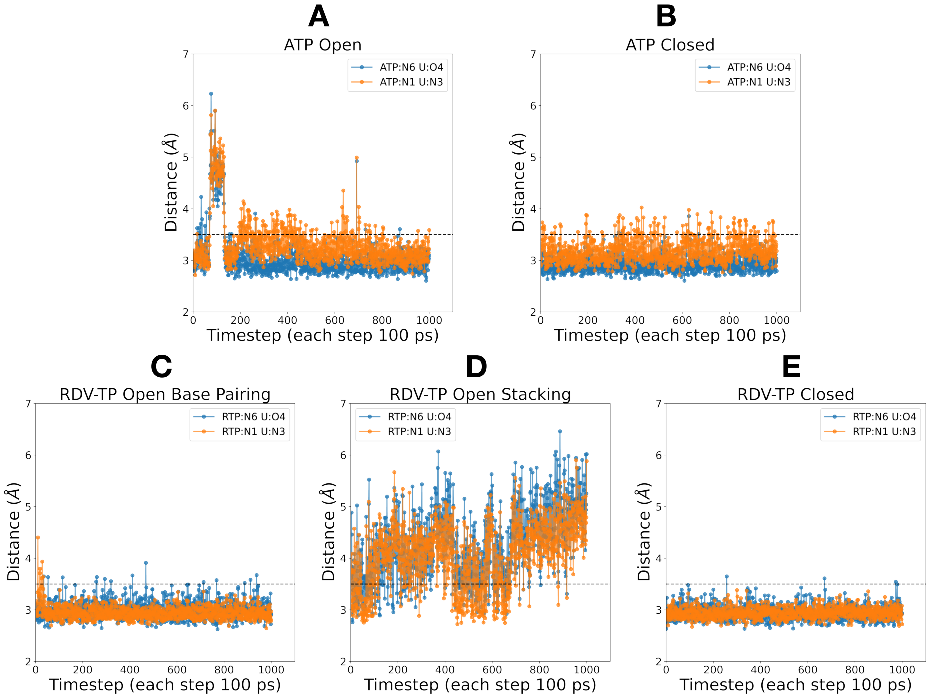


Figure 2.7: Expected hydrogen bond (HB) distance between ATP/RDV-TP and the +1 template Uracil (**A-E**). The distance pairs measured are the heavy atoms from the nucleotide triphosphate N6 with U:O4 and N1 and U:N3. The dashed black line indicate the cutoff (3.5 Å) for a HB. The NTPs in the insertion complexes (**B&E**) form significantly more stable HB than in in the initial binding forms. An RDV-TP in initial-binding forms a base stacking configuration with the template nt (**D**), in which hydrogen bonds are rarely formed.

much stabilized in the closed-state ATP insertion configuration (see Fig. 2.6A&B and Fig. 2.8A&B).

Upon MD equilibration of the open-state RdRp complex with RDV-TP ($\sim$ 100 ns; see Fig. 2.9 for RMSD), we found that RDV-TP shows primarily two unique open state binding configurations: one still with standard base pairing and the other with the RDV base stacking with the +1 template uracil base (see Fig. 2.6C&D).

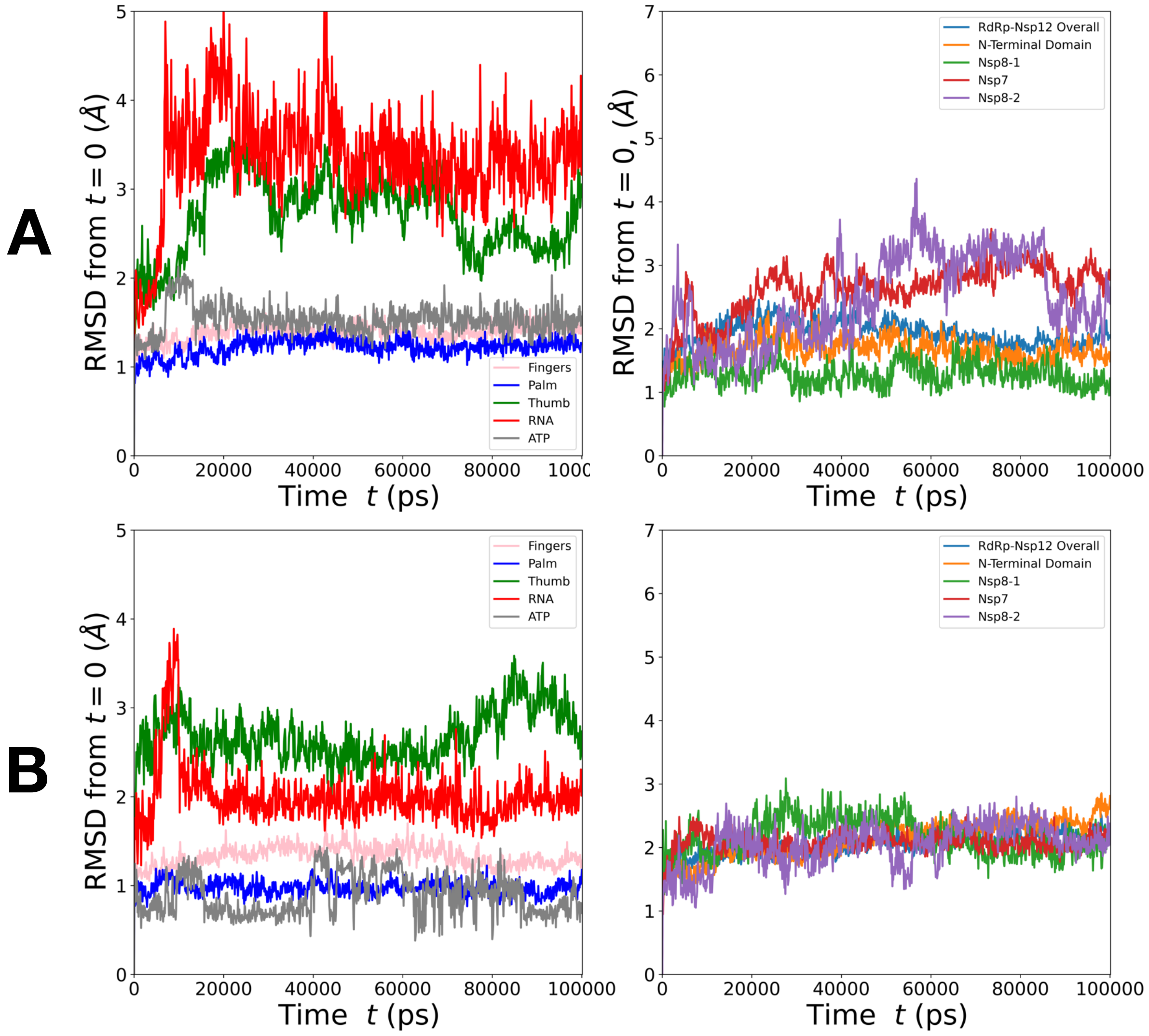


Figure 2.8: Measured RMSD from equilibration simulations. (Left) Subdomains, RNA, and ATP. (Right) Cofactors (ns7 and nsp8), Nsp12 and N-Terminus domain. **A** ATP initial binding complex. **B** ATP insertion complex. The insertion complex appears to be more stable than the initial binding complex.

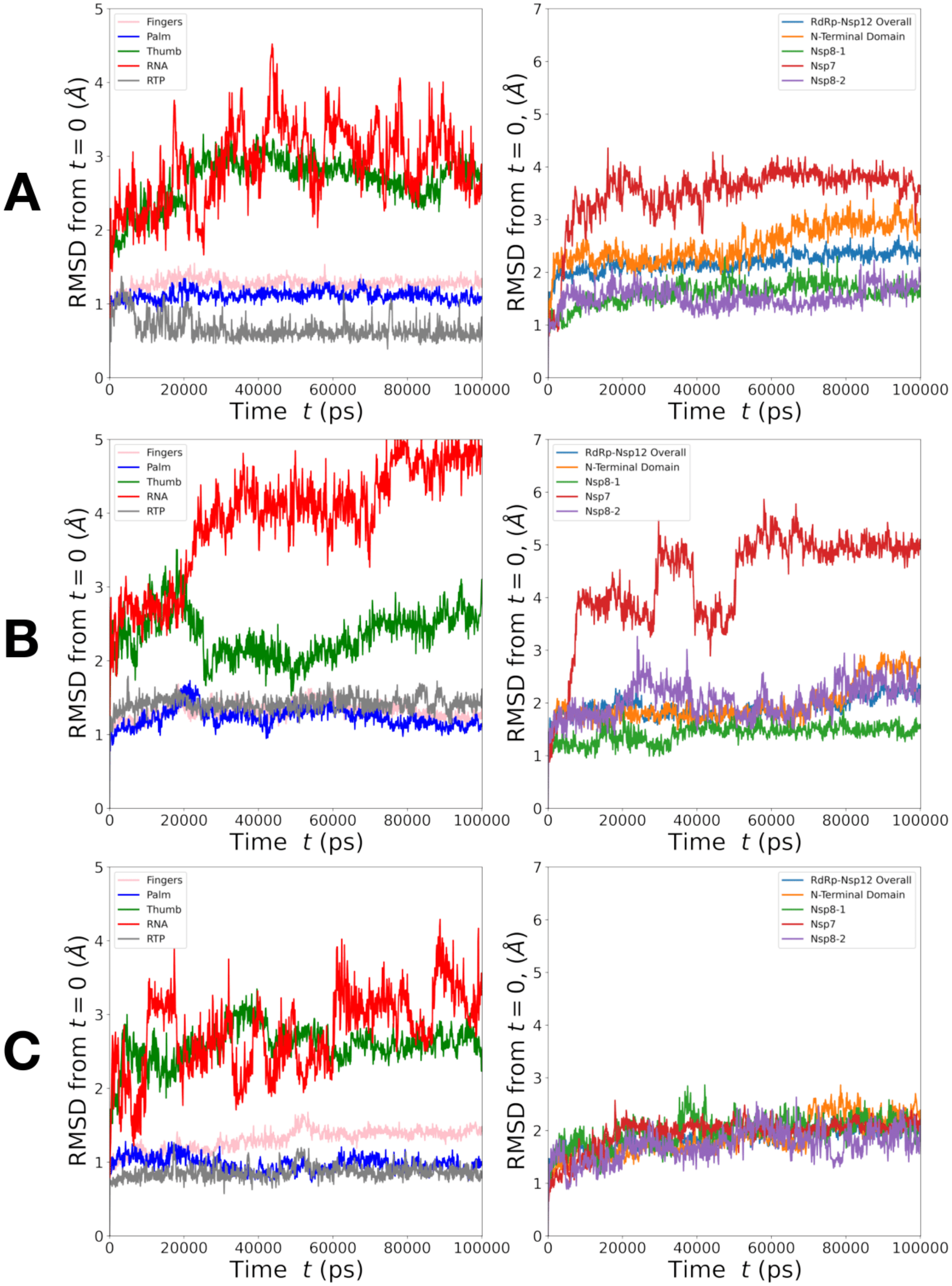


Figure 2.9: Measured RMSD from RTP (RDV-TP) equilibration simulations. (Left) Subdomains, RNA, and NTP. (Right) Cofactors (nsp and nsp8), Nsp12 and N-Terminus domain. **A** RDV-TP initial binding base pairing configuration. **B** RDV-TP initial binding stacking configuration. **C** RDV-TP insertion complex. The RDV-TP insertion complex appears to be more stable than the initial binding configurations.

2.4 Discussion

In this work we modeled and simulated insertion of the triphosphate form nucleotide analogue drug remdesivir (RDV-TP) into the SARS-CoV-2 RdRp active site, in comparison with natural nucleotide substrate ATP. Our work is based on high-resolution cryo-EM structures solved for the SARS-CoV-2 nsp12 in complex with cofactors nsp7 and nsp8,[7;11] modeled in an active-site open form (PDB: 7BTF) for the nucleotide initial binding, and in an active-site closed form (PDB: 7BV2) for the stabilized nucleotide insertion, prior to catalytic addition of the nucleotide to the synthesizing RNA chain. The viral RdRp or nsp12 in the coronavirus species works with other non-structural proteins (nsp7 to nsp16) for viral genome replication and transcription,[85;86] with nsp7 and nsp8 the cofactors to assist the replication machinery stability and processivity along the viral genome, and with nsp13[87;88] and nsp14[89] functioning as helicase and exonuclease, respectively. In the simulation of the nsp12-nsp7-nsp8 complex along with RNA strands, we found that shortening of the nsp8 N-terminal (e.g. to start from residue M67) is necessary to stabilize the simulation complex in all-atom explicit water condition. It is however noted that the two copies of nsp8 can extend very long as 'sliding poles' on a protruding exiting RNA duplex, as being captured from another high-resolution cryo-EM complex of nsp12-nsp7-nsp8.[68] In modeling of an initial binding complex of the nucleotide or analogue, we placed ATP or RDV-TP to the open active site of CoV-2 RdRp, according to RdRp structural alignments between the product complex (closed form) of RDV-TP and the open one. Accordingly, the positioning of RDV-TP or ATP appear similar between the open and closed structures. Molecular docking and simulation equilibration confirmed such an initial nucleotide binding configuration is dominant (see **SI Fig S3&S4**), which also shows similarly to that being captured in the poliovirus (PV) RdRp.[16] Hence, for the RDV-TP and ATP insertion probed in this work, we focus mainly on subtle local interactions around the active site of the viral RdRp as for the incoming nucleotide being recruited, interrogated, and re-positioned to allow chemical addition. Meanwhile, we note that the open and closed forms of the viral RdRp structure still involve collective movements of the highly conserved motifs (A to G) which we manipulate as a whole in the umbrella sampling simulations, to ensure the concerted nucleotide insertion. Note that motifs A to E are located in the palm subdomain hosting the active site, with motif C mainly responsible for catalysis, and motif A,B, and D for nucleotide binding and selection; motif F-G from in the fingers subdomain also impacts on the incoming nucleotide entry as well as the +1 template nt for the Watson-Crick (or WC) base pairing or fidelity check.[62;90]

Correspondingly, we conducted first the equilibrium MD simulations, which show that upon the initial binding, ATP frequently forms the WC base pairing with the template nt but with notable fluctuations; in contrast, RDV-TP primarily forms base stacking with the template nt, squeezing the template base to upstream most of time. Although RDV-TP has also been sampled in base paring with the template uracil base, such a base stacking configuration appears more stable. In the closed RdRp or insertion state, RDV-TP anyhow forms highly stabilized base pairing with the template nt, with even lower fluctuations than ATP for natural base pairing. APBS mapping zoomed into the closed active site of CoV-2 RdRp shows notable differences between the local electrostatic environment around the inserted RDV-TP and ATP (see **Fig S2**), in particular around the sugar region, where an extra cyano group is attached to RDV-TP, with T687 and N691 associated nearby. In order to see how exactly RDV-TP and ATP insert into the active site from the initial binding state, as the open active site closes, we then performed the TMD and umbrella sampling simulations connecting the open and closed RdRp complex structures, with slightly varied initial and collective coordinate forcing conditions.

Chapter 3

Thermodynamic Quantification of Nucleotide Insertion: A Free Energy Analysis of ATP and Remdesivir

In this chapter, we undertake a comparative analysis of the computed free energy of insertion between ATP and Remdesivir-TP. To achieve this, we employ two protocols: one where force is applied to the template nucleotide at position +1, and another where no force is applied within our defined reaction coordinate. For Remdesivir, we also consider two initial binding or open configurations - one that involves Watson-Crick base pairing and another where it stacks with the uracil template. We continue with the models constructed and equilibrated as discussed in Chapter 2. Through the computation of the free energy of insertion, we demonstrate that Remdesivir, particularly in the stacking configuration, inserts more readily into the closed state than ATP, although the overall ΔG of the two compounds remains relatively similar.

3.1 Introduction

It's important to recognize that in single-subunit viral RNAPs or DNAPs, the nucleotide insertion, in accompany with the open to closed conformational transition (pre-chemistry transition or isomerization), usually happens slowly (e.g. milliseconds or above), i.e., to be rate limiting (or partially rate-limiting) in the NAC.[91–93] Such a slow nucleotide insertion step correspondingly plays a significant role in the nucleotide selection or fidelity control, for example, in the single-subunit viral T7 RNAP system studied recently.[94–96] To understand how RDV-TP can evade from the nucleotide selectivity of RdRp to be incorporated, it is therefore essential to probe how such a nucleotide analogue binds stably and inserts sufficiently fast or with low energy barrier into the active site, comparing to its natural substrate counterpart. Accordingly, in this work, we employed all-atom MD simulation to probe mainly the free energetics of the RDV-TP insertion into the CoV-2-RdRp active site, in comparison with the ATP insertion. To do that, umbrella sampling strategies were implemented connecting the initial substrate binding (active site open) and the insertion (active site closed) conformational states, in particular, by enforcing collective coordinates of atoms from structural motif A-G and the inserting NTP (excluding or including the template nucleotide or nt +1 with forcing). The simulations consequently reveal free energetics or potentials of mean force (PMFs) along the reaction coordinate of the RDV-TP and ATP insertion, demonstrating how local residues around the RdRp active site or NTP binding site coordinate with the nucleotide binding, insertion, and differentiation, comparing RDV-TP and ATP.

3.2 Determining the Reaction Coordinate and Calculating Free Energy

The open to close conformational change of the RdRp is expected to be on the order of milliseconds and therefore can not be captured by brute force MD. In order to calculate free energy, the umbrella sampling method was used.[45;49;97] To use such a method a reaction path needs to be specified and followed. In this study we used TMD to generate such a path between the open and closed states. TMD[98] implementation requires an initial and a final reference structures to be specified which we continue using in the umbrella sampling simulations. In this work we implemented two slightly varied protocols by manipulating coordinates of two slightly varied atom sets: nsp12 motifs (motif A-G backbone atoms) and NTP (heavy atoms), with or without template +1 nt (heavy atoms). The corresponding RC is then constructed by aligning the structures to the reference structures via the fingers sub-domain and measuring the differences of RMSDs.

$$RC(X) = \delta RMSD(X) = RMSD(X, X_{\text{Open ref}}) - RMSD(X, X_{\text{Closed ref}}) \tag{3.1}$$

Where X is the coordinates for the above selected atom sets and $X_{\text{Open ref}}/X_{\text{Closed ref}}$ is for a chosen reference state.

3.2.1 Selecting Reference Structures

The reference states used for the reaction coordinate or the implementation of TMD need to be close to equilibrium but not at equilibrium, since we want to sample both sides of equilibrium region along the RC, while the reference structures correspond to the two ends of the RC$(-RC_{\text{max}}, +RC_{\text{max}})$, with $RC_{\text{max}} = \delta RMSD(X_{\text{Open ref}}, X_{\text{Closed ref}})$. The reference structures or states are selected using the first 50ns of the unrestrained NPT simulations, and the correspondingly defined RCs for the open and closed equilibrated structures need to satisfy the conditions below:

$$\begin{aligned}\delta RMSD(X_{\text{Open equi}}) =& RMSD(X_{\text{Open equi}}, X_{\text{Open ref}}) - \\ & RMSD(X_{\text{Open equi}}, X_{\text{Closed ref}}) \\ \delta RMSD(r_{\text{Closed equi}}) =& RMSD(X_{\text{Closed equi}}, X_{\text{Open ref}}) - \\ & RMSD(X_{\text{Closed equi}}, X_{\text{Closed ref}})\end{aligned} \tag{3.2}$$

Where the requirement is:

$$\begin{aligned}-RC_{max} < RC(X_{\text{Open equi}}) < 0 & \\ & 0 < RC(X_{\text{Closed equi}}) < +RC_{max}\end{aligned} \tag{3.3}$$

Where the RC is $\delta RMSD$ specified in Eq. (3.1).

3.2.2 Target MD and Umbrella Sampling

Using the selected open and closed reference structures, the TMD is launched from each state to create paths (forward path started from the open to the closed reference structure, and the backward path started from the closed then to the open reference structure) that meet halfway on the RC (see Fig. 3.1). From the forward and backward TMD paths created

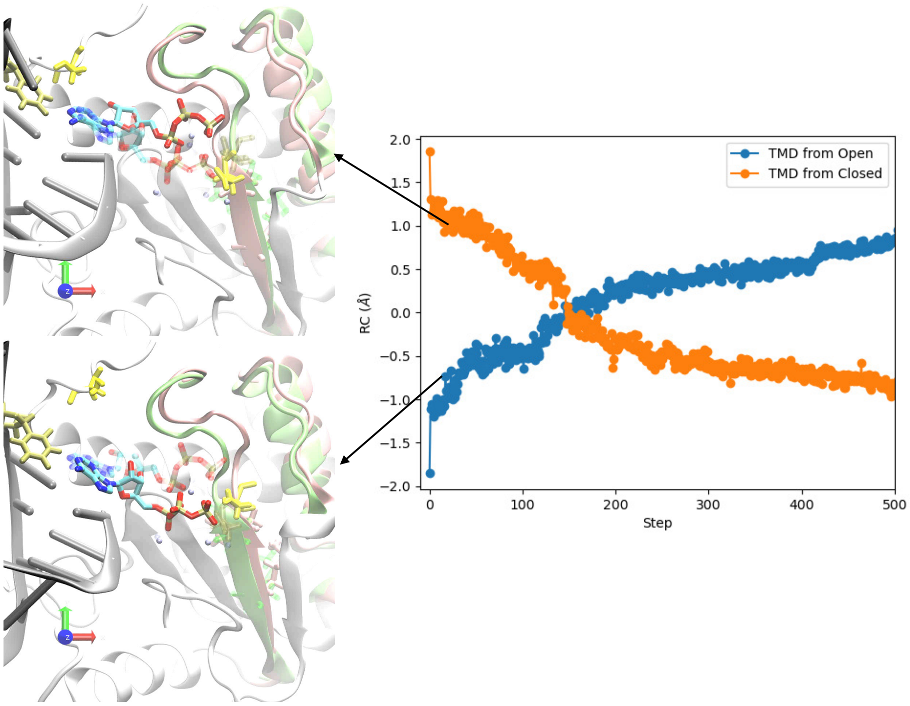


Figure 3.1: Implementation of targeted molecular dynamics (TMD) simulations for constructing NTP insertion path to be utilized in the umbrella sampling simulations. (Left) The initial and final structures of each respective paths: backward (bottom) and forward (upper). With motifs A/D colored pink for the starting structure and green for the target structure. Representations are colored to compare with PV RdRp in.[13] (Right) The implementation of the TMD simulations forward and backward. Where TMD from open refers to starting from the initial binding complex (forward path), and TMD from closed is starting from the insertion complex (backward path). Structures are selected every 0.1 Å from the two paths until they meet in the middle along the reaction coordinate (RC). Such a constructed path is then used for the umbrella sampling simulation.

between the open and closed states, structures are evenly (for every 0.1 Angstrom in the RC) selected to be used for umbrella sampling simulations. In the umbrella sampling simulations from the selected structures along the TMD paths, harmonic restraints are used along the RC. The force constants used in TMD are subsequently used in the umbrella sampling simulations (see Table. 3.1).

Table 3.1: Summary of target MD and Umbrella Sampling parameters. The force constant used from the TMD simulations were carried over and used for the respective umbrella sampling simulations. Large force constants were used for the ATP simulations and smaller ones for RDV-TP simulations. Where the () in RDV-TP or RTP systems indicate the initial binding structure (open for the active site open state).

RC	Force Constant $\left(\frac{kcal}{mol\mathring{A}^2}\right)$	RC Range(Å)	Number of Windows
Motifs + ATP	501	-1.2 to 1.3	27
Motifs + ATP + Template	501	-1.1 to 1.3	26
Motifs + RTP(Open Stacking)	125	-1.0 to 1.0	21
Motifs + RTP(Open Stacking) + Template	125	-1.6 to 1.6	34
Motifs + RTP(Open Base-pairing) + Template	250	-1.5 to 1.5	32

The biased histograms along the RC for each window were unbiased / re-weighted using the weighted histogram analysis method.[99] From the generated biased trajectories a short set of data is removed from the beginning of each for equilibration (10 ns for RTP simulations and 20 ns for ATP simulations as it takes longer time for ATP simulation systems to converge). Overlap for each set of windows was checked along the reaction coordinate (see Fig. 3.2).

The unbiased probabilities and then the free energy are also calculated using WHAM package,[101] following equations:

$$\begin{aligned} P_i(RC) &= \exp\left[\frac{-k(RC - RC^o)^2}{2k_BT}\right] P_i^{'}(RC) \\ G(RC) &= -k_BT \ln P_i(RC) \end{aligned} \tag{3.4}$$

Where $P_i(RC)$ and $P_i^{'}(RC)$ are the unbiased and biased probabilities sampled for the i-th window, respectively. The harmonic restraint potential is shown by $\frac{k}{2}(RC - RC^o)^2$ where RC^o is for the initial structure obtained from the TMD insertion path. Finally free energy profile G along the RC is calculated taking the logarithm of unbiased probabilities, which represent the PMF.

While constructing the PMF using WHAM, bootstrapping error analysis[102] is used to

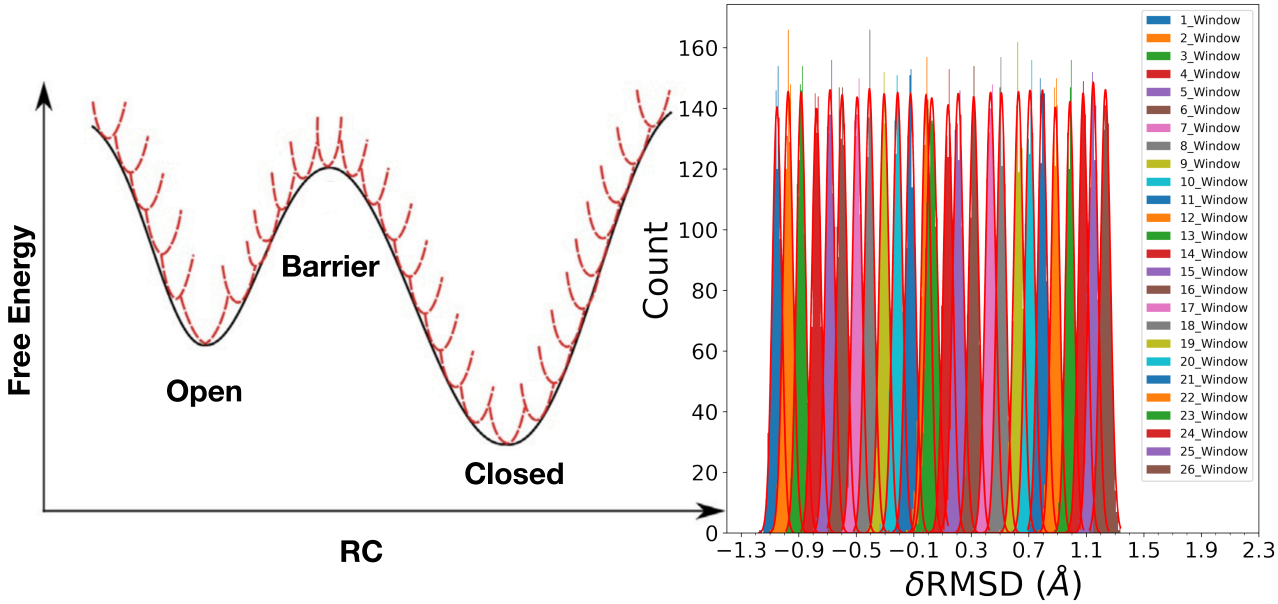


Figure 3.2: Conducting the umbrella sampling simulations for the NTP insertion. (Left) The schematics of the umbrella sampling simulation strategies (figure adapted from[100]). (Right) The overlap of simulated windows, where the RC is centered every 0.1Å, with the initial simulation structure taken from the forward/backward TMD paths (see Fig. 3.1).

estimate errors. Bootstrapping re-samples RC_i in each window; from each bootstrapped trajectory $RC_{b,i}(t)$ a new histogram ($h_{b,i}(RC)$) is created. From the new histograms the PMF and $G_b(RC)$ are reconstructed, this process is repeated N times (N = 500 used in this study) generating N bootstrapped PMFs $G_{b,j}(RC)(j = 1, 2, ..., N)$. The uncertainty of a PMF is estimated by a standard deviation calculated by the N bootstrapped PMF's.

$$\sigma_{PMF}(RC) = \left[(N-1)^{-1} \sum_{j=1}^{N} (G_{b,j}(RC) - \langle G_b(RC) \rangle)^2 \right]^{\frac{1}{2}} \tag{3.5}$$

Example input files for both TMD and U.S. are included in the 6.2

3.2.3 Hydrogen Bond Analysis around NTP

To examine the corresponding nucleotide insertion dynamics (with intermediate or transition state over-sampled in the umbrella sampling simulations), hydrogen bond (HB) analysis was performed on the trajectories sampled along the RC of the NTP insertion from open to closed. This was done by taking the last 10 ns of each window and combining them into a single trajectory for each simulation system. HBs were measured using the MDAnalysis[103] python package with a heavy atom donor-acceptor distance cutoff of 3.5 Å and angle cutoff of 140°. From this analysis, plots were created to indicate when a particular HB was present or not from open to close along the RC. Using a similar strategy plots for electrostatic interaction (with a distance cutoff 5 Å) of salt bridges for the NTP polyphosphate were also constructed (see Fig. 3.3). Stacking was determined by measuring whether the two base rings are parallel and overlap well.(see Fig. 3.4). The Mg^{2+} ions were also analyzed by measuring the distances between the Mg^{2+} ion and the NTP center of geometry (C.O.G.) (see Fig. 3.5). The measurements show that only the two of the three Mg^{2+} ions (MgA and MgB) are comparatively stabilized near the bound NTP substrate, while the third one (MgC) stays a bit far, suggesting that the third Mg^{2+} does not play as much of a role in coordination as the two catalytically important Mg^{2+} ions.

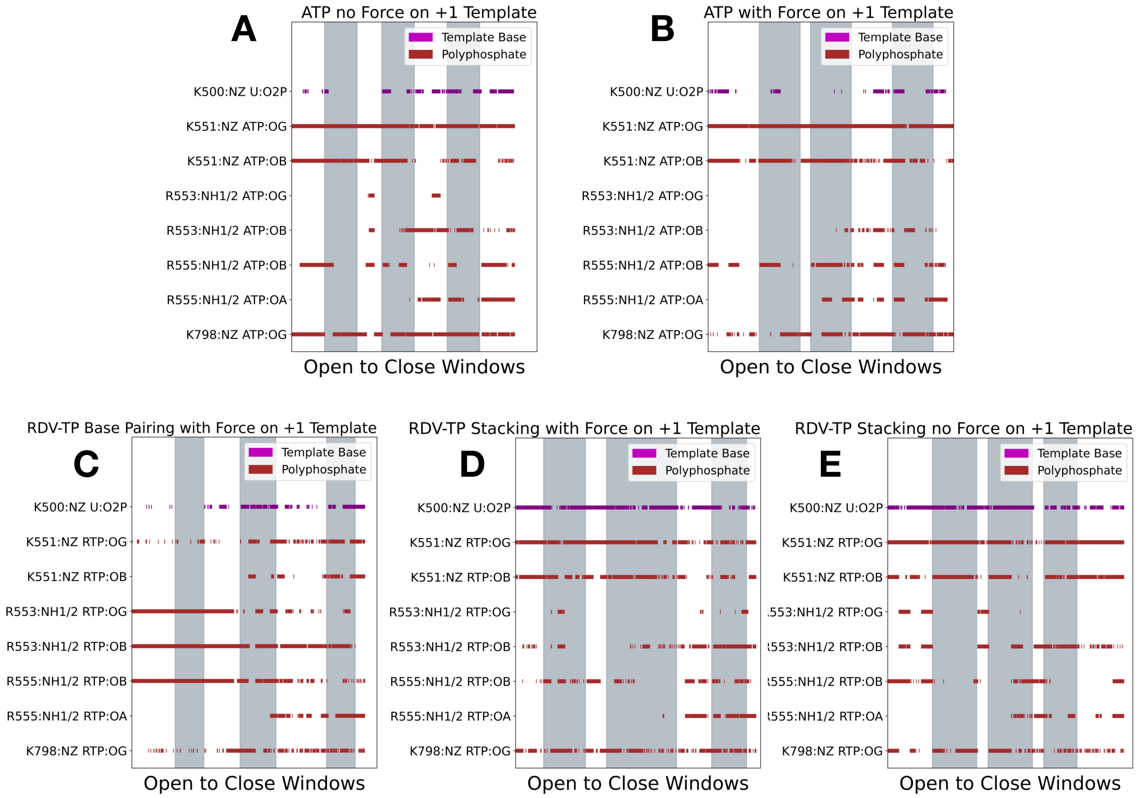


Figure 3.3: Salt-bridge electrostatic interactions with ATP/RDV-TP triphosphates. Here we can identify the positively charged residues (Lys and Arg) which can form salt bridges with the negatively charged oxygen's along the polyphosphate. Distances are measured from the positive charge center (NZ nitrogen in Lys and CZ carbon in Arg) and the negative charge (O1G,O2G,O3G,O1B,O2B,O1A, and O2A in the NTP or O1P and O2P in the template backbone), if the charges are less than 5Å[104] a salt bridge is identified.

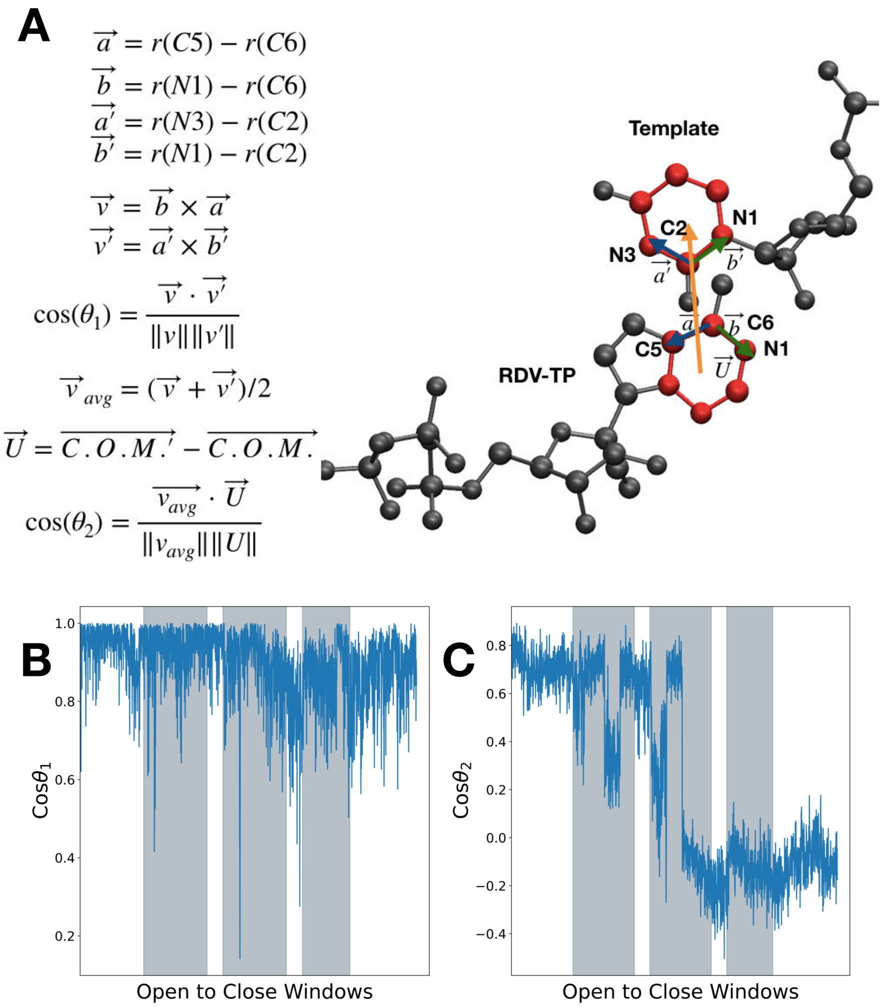


Figure 3.4: Schematics for base-stacking measurements. **A** The stacking is determined by measuring if the two base rings are parallel with reasonable overlapping. **B** The first $\cos\theta_1$ is calculated from atoms within the six numbered rings in the bases. **C** The second $\cos\theta_2$ is calculated via considering the center of mass (C.O.M.) of the two six numbered rings. Reasonable base stacking is formed when $\cos\theta_1 > 0.8$ (θ_1 close to zero or the two rings being parallel) and $\cos\theta_2 > 0.6$ (θ_2 not far from zero or the two rings overlap). Measurements shown are from the RDV-TP with base stacking (without force on template nt) insertion.

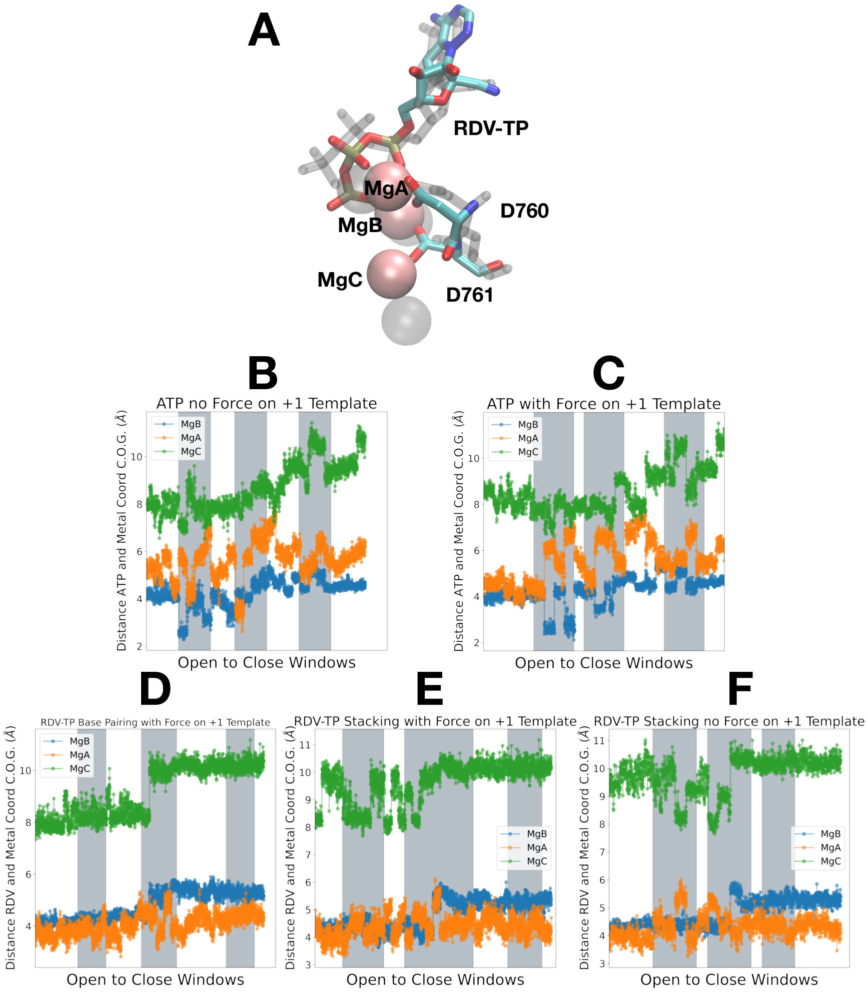


Figure 3.5: **A** The modeled insertion state structure of RDV-TP (grey transparent) aligned with that of the equilibrated insertion state one (in color). MgA is in coordination with residues D760 and D761. MgB is coordinated by the β and γ phosphate oxygens. **B-F** Distances measured between the center of geometry (C.O.G.) of the NTP and Mg^{2+} ions in the umbrella sampling simulations (from open to barrier and to the closed state). MgA is in coordination with the catalytic D760 and D761 residues as well as the 3' end primer backbone phosphate. MgB is in coordination with the phosphate group (β & γ phosphate oxygens).[19] MgC is unlikely to be involved in catalysis or product release as it stays comparatively far from the NTP.

3.3 Free Energies of Insertion for ATP and RDV-TP

We performed TMD simulations to generate the nucleotide substrate (ATP and RDV-TP) insertion paths, and finally conducted a series of umbrella sampling simulations to obtain the nucleotide insertion PMFs for individual systems. The results show uniformly that the closed insertion state is more stabilized than the open initial binding state for each substrate, while the relative stability of the open states ($\Delta G^{OC} = G_{Open} - G_{Closed}$) and the insertion barriers ($\Delta h^{ins} = G^{Barrier} - G_{Open}$) vary for individual systems. We illustrate results on these systems below, for the ATP insertion, (i) excluding and (ii) including +1 template nt in the RC, initiated from the open state, with ATP base pairing with the template nt; for the RDV-TP insertion, initiated similarly from the (iii) RDV-TP base pairing with the template nt under forcing (i.e., included in the RC), and then from a varied initial configuration, i.e., RDV-TP stacking with the template nt, as the nt (iv) included and (iv) excluded in the RC (i.e., with and without forcing).

3.3.1 Insertion of ATP into the active site can be facilitated by base pairing with the stabilized template nt (+1)

By obtaining quasi-equilibrated reference structures from the open-state ATP binding and closed-state ATP insertion complexes, we performed the TMD simulation between these two reference structures and constructed the ATP insertion path for conducting the umbrella sampling simulations. The convergence of the PMFs for the ATP insertion requires about 100$\sim$200 ns MD simulation for individual simulation window (see Fig. 3.6A&B). In the first simulation system, ATP constantly forms base pairing with the +1 template nt in the initial binding or active-site open state. We conducted the umbrella sampling simulations by forcing atoms from motif A-G and ATP along the TMD insertion path. In this case, the +1 template nt is excluded from the RC, so it is subject only to thermal fluctuations but not the umbrella forcing or constraining. Under such conditions, we noticed that ATP can become highly destabilized by occasionally shifting its base far from the active site in the open state and during barrier crossing (see Fig. 3.7). Overall, the ATP insertion can still proceed toward the comparatively stabilized closed state, with ATP base pairing with the template nt much better than in the open state. Correspondingly, the open to closed free

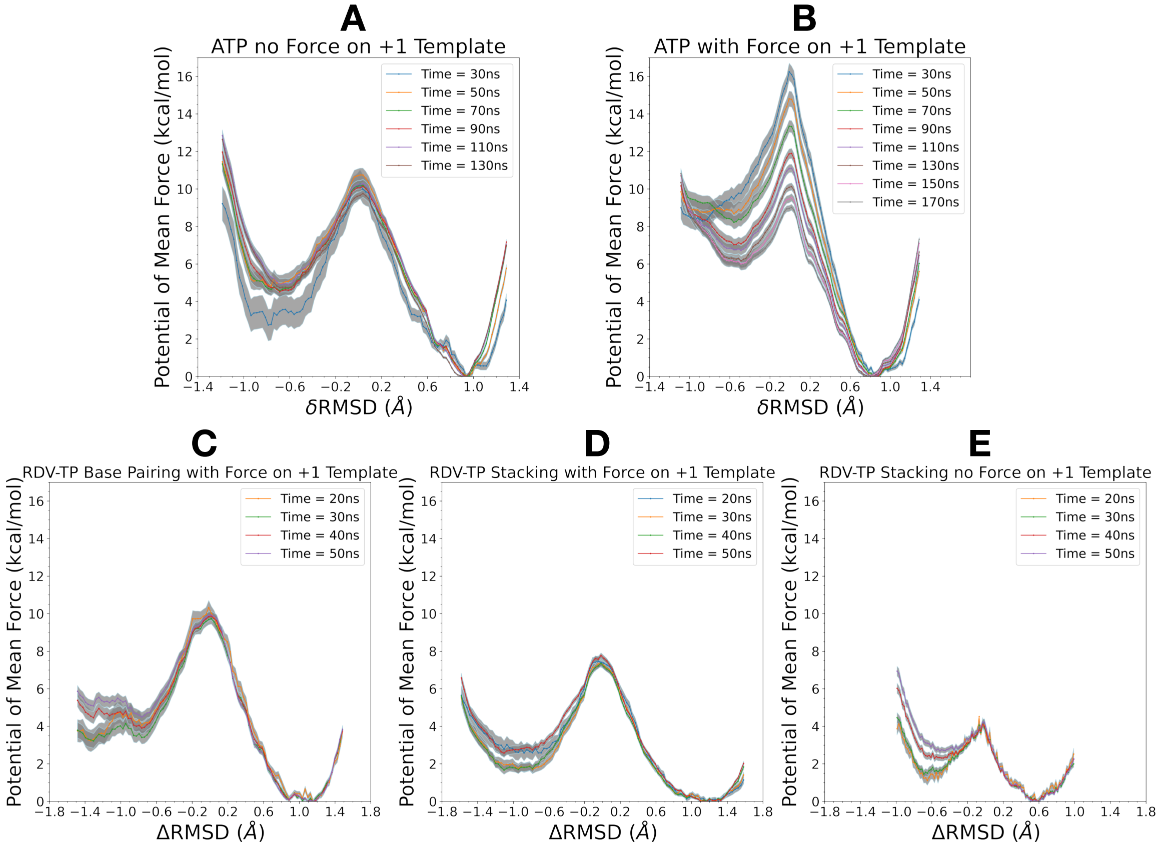


Figure 3.6: Convergence plots of all PMFs constructed with bootstrapping error analysis for each set of data.[96;102] As more data are accumulated with the extended simulations, the PMF further converges. Early data collected is removed as for pre-equilibration, 20ns for ATP and 10ns for RDV-TP systems. **A&B** ATP PMF's with no force on template and with force on template, respectively. **C** RDV-TP base pairing with force on +1 template nt. **D** RDV-TP stacking with force on + 1 template nt. **E** RDV-TP stacking with no force on + 1 template. **D-E** Only 50ns of data from each window was needed to reach convergence.

energy drop is obtained as ΔG^{OC} ∼4.8±0.3 kcal/mol and the ATP insertion barriers appears high as Δh^{ins} ∼5.0±0.3 kcal/mol. During insertion, one can see that motif F-K551 (R555) and K798 (near motif D C-term) constantly form HB interactions with the triphosphate of ATP throughout the process, along with motif F-K545 and the template nt; motif C-D760 form occasional HBs with the ATP sugar at open state to barrier crossing, but not into the closed state; motif B-N691 and motif A-D623 form no HBs with ATP sugar until the closed state or crossing the barrier, along with motif F-R553 with the ATP phosphate and motif

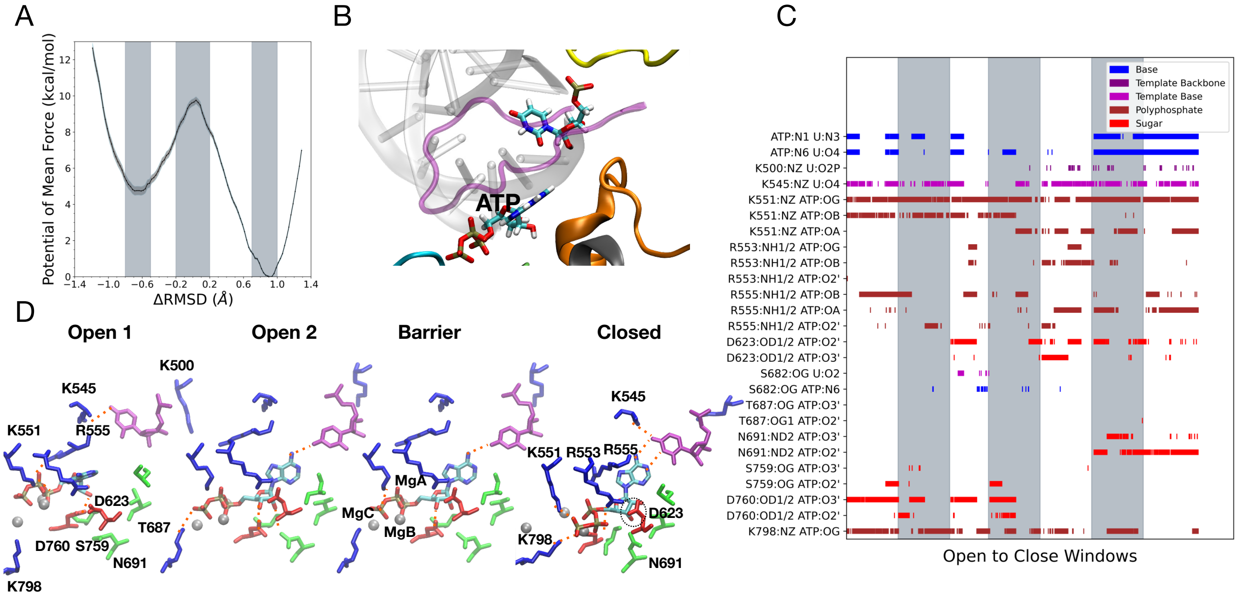


Figure 3.7: ATP insertion from umbrella sampling MD simulation (without force on the +1 template nt in the RC). **A** PMF with barrier 5.0±0.3 kcal/mol and an initial binding stability of 4.8±0.3 kcal/mol. **B** Open conformation with ATP not forming hydrogen bonds with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows (see Fig. 3.3A for salt bridges). **D** Interaction snapshots from simulation windows: Two open states are shown due to the volatility of the open minima, ATP often flips out of plane from the +1 template base. As the barrier is crossed it begins to form consistent base pairing with the template. Dotted orange lines highlight essential HB interactions.

G-K500 with the template backbone.

Next, in order to stabilize the ATP insertion process, we included the +1 template nt in the RC (i.e., with the umbrella forcing) and constructed the second PMF (see Fig. 3.8). Consequently, with the ATP:template nt base pairing is better stabilized. The ATP base deviated less frequently and not that far from the active site in the open to the barrier crossing state, and ATP base pairing with the template nt can recover sooner after barrier crossing. Notably, the ATP insertion barrier lowers to Δh^{ins} ~2.6±0.3 kcal/mol, although the initial open state stability maintains similarly as in the first case (or slightly less stabilized: ΔG^{OC} ~5.1±0.2 kcal/mol relative to the closed state). Hence, forcing on the template nt or quenching the fluctuations seems to facilitate the ATP insertion, likely by stabilizing the transition state with the ATP-template nt base paring. Such an operation can mimic the

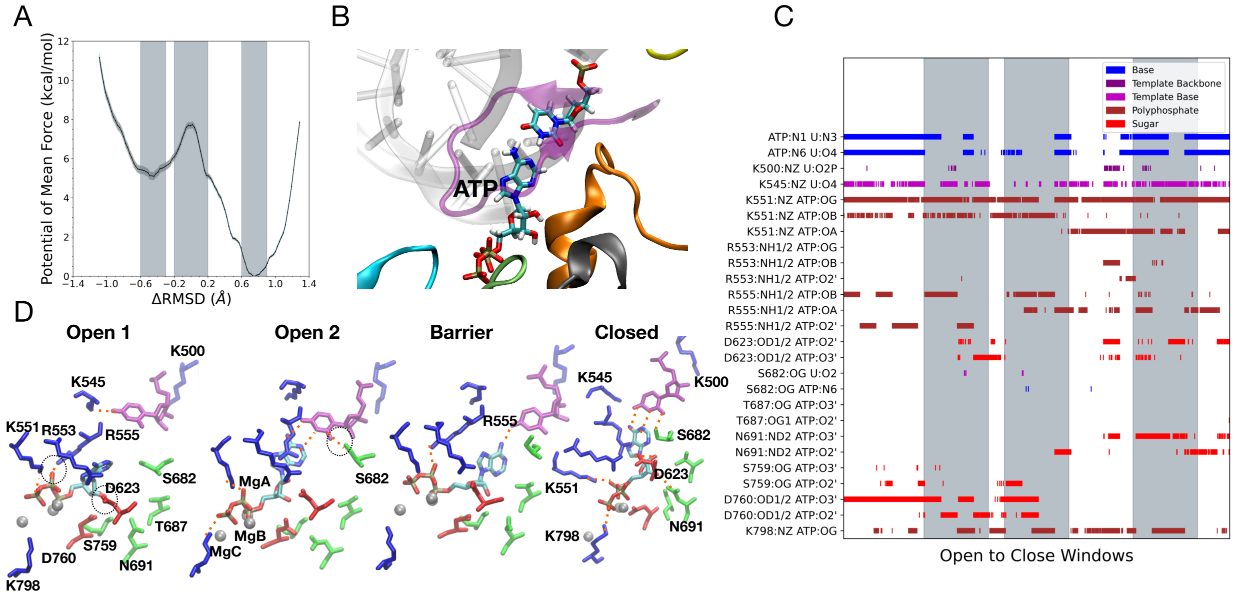


Figure 3.8: ATP insertion from umbrella sampling MD simulation (with force on the +1 template nt in the RC). **A** PMF with barrier of 2.6 ± 0.3 kcal/mol and an initial binding stability of 5.1±0.2 kcal/mol. **B** Open conformation of ATP forming hydrogen bonds with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3B for salt bridges). **D** Interaction snapshots from simulation windows: Two open states are shown due to the volatility of the open minima. Although ATP still occasionally flips out of plane, it more consistently forms HB with the +1 template base. Dotted orange lines highlight essential HB interactions.

spontaneous ATP insertion process that happens sufficiently slowly (e.g. over milliseconds). Overall, the ATP local interactions with nearby amino acids around the active site appear similarly in the two simulation systems, except that in the current template forced condition, the HBs from motif A-D623:sugar and motif G-K500: template formed a bit earlier in the open state, and motif F-R555 forms HBs with the ATP phosphates more often throughout the process. Hence, the D623-sugar, R555-phosphate, and the K500 template nt interactions, along with the template forcing on stabilizing the ATP-template nt base pairing seem to contribute to the lowered ATP insertion barrier.

3.3.2 RDV-TP initial stacking with the +1 template nt is more stabilized than the base pairing

We next constructed the PMF for the RDV-TP initially base pairing with the template nt (see Fig. 3.9), applying force or constraint to the template (similarly as to the ATP insertion in Fig. 3.8). Then we chose the varied initial binding configuration as the RDV forms base stacking with the template nt, keeping the force constraint on the template, and repeated the calculations (see Fig. 3.10). Note that the convergences of the RDV-TP insertion energetics happen much faster (∼50 ns; see Fig. 3.6C-E) than that of the ATP system. The PMF of the RDV-TP insertion starting from the base pairing configuration

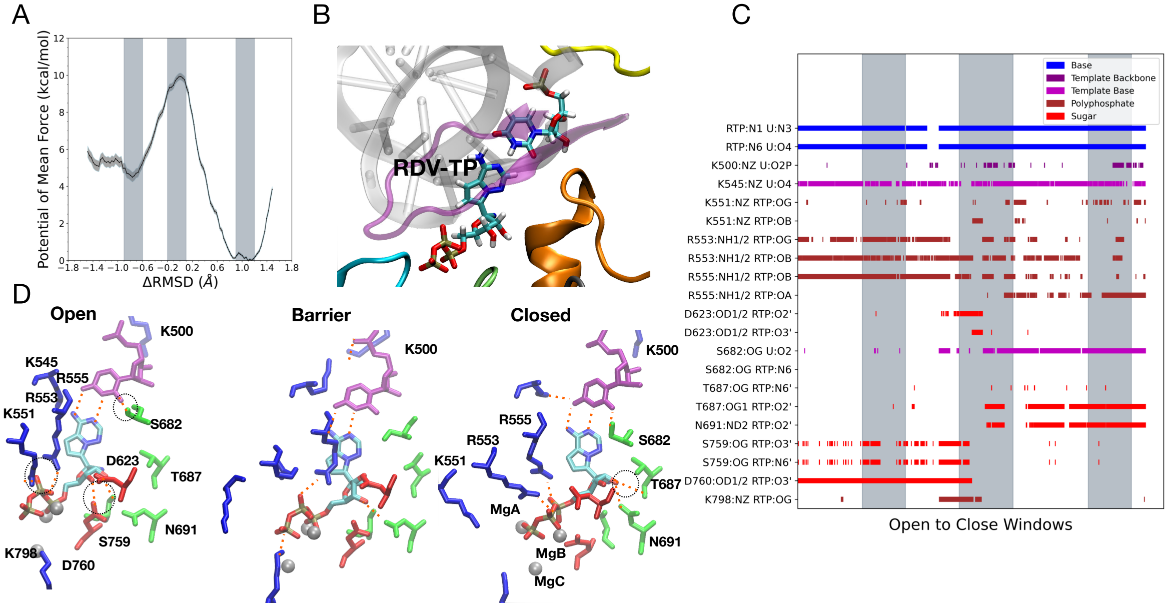


Figure 3.9: RDV-TP insertion with the open state forming good base pairing from umbrella sampling MD simulation (with force on the +1 template nt in the RC). **A** PMF with barrier of 5.4 ± 0.3 kcal/mol and an initial binding stability of 2.6±0.3 kcal/mol. **B** Open conformation of RDV-TP forming hydrogen bonds with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3C for salt bridges). **D** Interaction snapshots from simulation windows: Throughout the open state stable HB form with the RDV and +1 template base. Dotted orange lines highlight essential HB interactions.

shows that the insertion barrier is high (Δh^{ins} ∼5.4±0.3 kcal/mol), comparing to the ATP insertion barrier obtained in the similar conditions (Δh^{ins} ∼2.6±0.3 kcal/mol from Fig. 3.8).

The relative stability of the open initial binding state of RDV-TP to the closed insertion state is also measured (ΔG^{OC} ~4.5±0.3 kcal/mol), slightly more stabilized, relatively, than that in the corresponding ATP base pairing system (ΔG^{OC} ~5.1±0.2 kcal/mol from Fig. 3.8). Now motif F-K551, R553&R555 form HBs less or more with the triphosphate of RDV-TP throughout the process, along with motif F-K545 with the template; motif C-S759 and D760 form HBs with the sugar at open state to barrier crossing, not afterwards or into the closed state; motif B-N691 and motif B-T687 barely forms HB with the sugar until the barrier crossing, along with motif G-K500 and the template. Overall, motif F-R553 and R555 form stronger interaction with the RDV-TP triphosphate than in the ATP insertion cases, while motif A-D623 barely forms HB with the RDV-TP sugar into the closed or insertion state (but with ATP sugar in the insertion state). In contrast, motif B-T687 forms HB with the RDV-TP sugar in the insertion state, while there is no HB interaction of it with ATP at all.

More interesting results come from comparing RDV-TP insertion energetics and interactions simulated at the varied conditions. In Fig. 3.10, we show the PMF from RDV-TP initially stacking with the +1 template nt, with forcing still implemented. Although the insertion barrier (Δh^{ins} ~5.2±0.3 kcal/mol) remains similarly high as the above case (from Fig. 3.9), the relative stability of the initial open state to the final insertion or closed state changes (to ΔG^{OC} ~2.6±0.2 kcal/mol), indicating that the initial stacking configuration of RDV-TP is more stabilized (about -3 k_BT) than the initial base pairing configuration with the template nt). By comparing the HB patterns (Fig. 3.9C and Fig. 3.10C), one finds that the stabilizing interactions to the base stacking configuration at the open state mainly come from motif A-D623 and motif C-S759 with the sugar, S682 interaction with the RDV-TP base, K798 (near motif D) along with motif F-K551 interaction with the phosphate, as well as motif G-K500 interaction with the template nt. The motif F R555/R553 interaction with the RDV-TP triphosphate weaken from the initial base pairing to the stacking configuration.

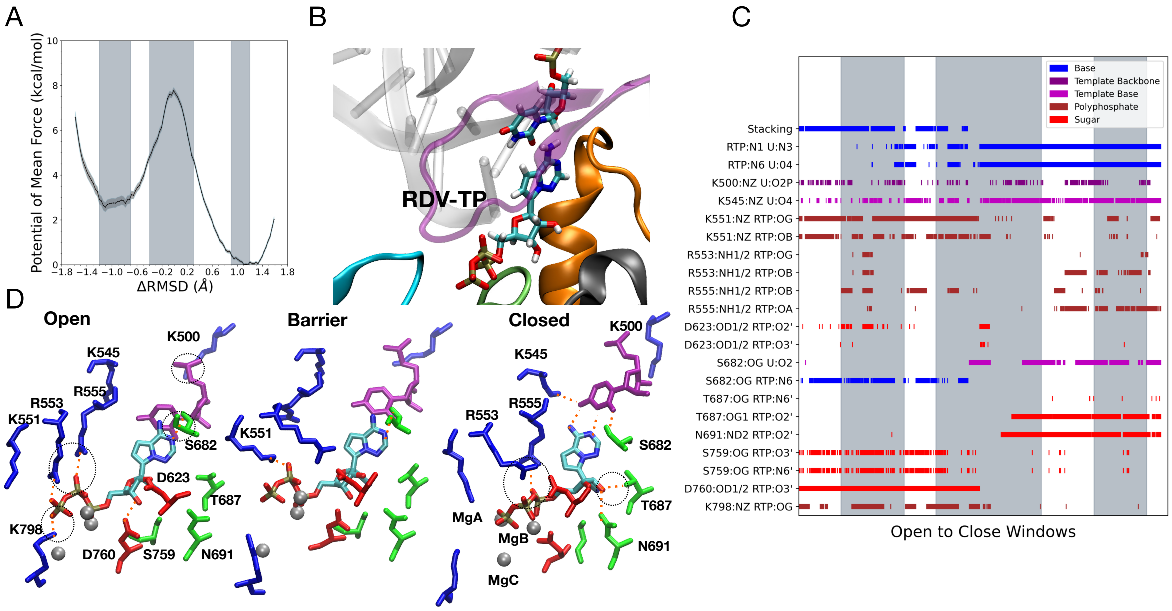


Figure 3.10: RDV-TP insertion with the open state forming base stacking with the +1 template base from umbrella sampling MD simulation (with force on the +1 template nt in the RC). **A** PMF with barrier of 5.2 ± 0.3 kcal/mol and an initial binding stability of 2.6±0.2 kcal/mol. **B** Open conformation of RDV-TP forming base stacking with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3D for salt bridges). **D** Interaction snapshots from simulation windows: Throughout the open state base stacking forms resulting in a more stable minima. Dotted orange lines highlight essential HB interactions.

3.3.3 RDV-TP insertion to the active site is facilitated by thermal fluctuations and absence of motif F-R553/R555 interaction with the polyphosphate

Since the above results show that the RDV-TP initial stacking with the +1 template nt is more stabilized than the base pairing configuration, we further explored the RDV-TP insertion barrier by removing the forcing on the +1 template nt (i.e. being excluded from the RC). Notably, the insertion now is greatly facilitated by allowing sufficient fluctuations on the template , such that the insertion barrier becomes lowest (Δh^{ins} ∼1.5±0.2 kcal/mol; see Fig. 3.11A). Meanwhile, the relative stability of the open binding state to the closed insertion state of RDV-TP maintains (ΔG^{OC} ∼2.7±0.1 kcal/mol), as in the above system from Fig. 3.10).

It appears that thermal fluctuations on the template nt can actually support the RDV

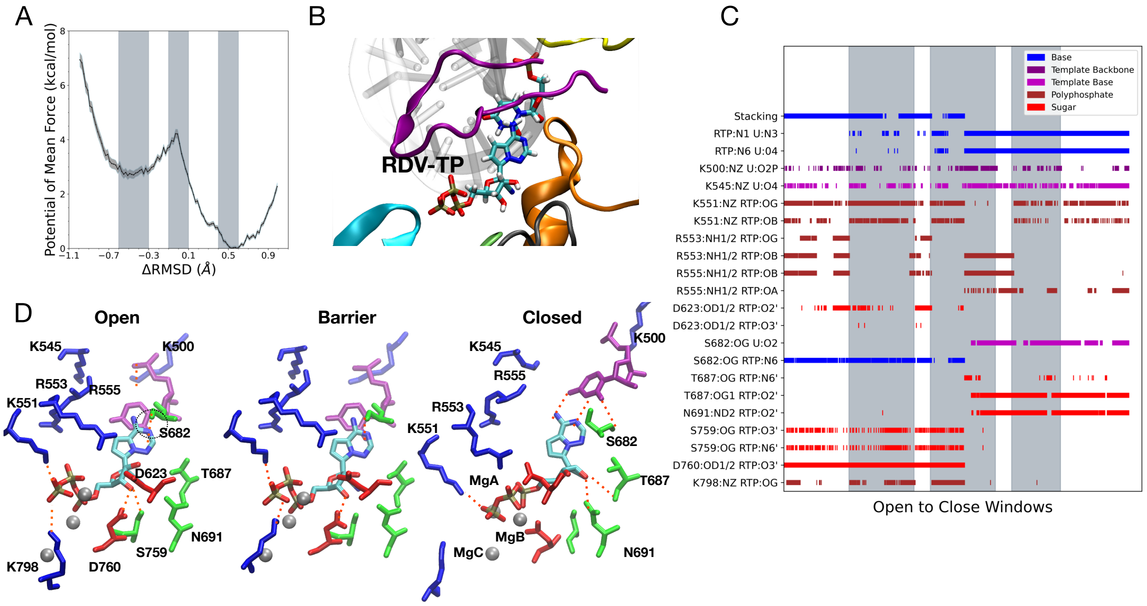


Figure 3.11: RDV-TP insertion with the open state forming base stacking with the +1 template base from umbrella sampling MD simulation (with no force on the +1 template nt in the RC). **A** PMF with barrier of 1.5 ± 0.2 kcal/mol and an initial binding stability of 2.7±0.1 kcal/mol. **B** Open conformation of RDV-TP forming base stacking with +1 template base. **C** Systematical HB patterns; the grey bars represent Open, Barrier, and Closed regions of the simulation windows as shown in the PMF (see Fig. 3.3E for salt bridges). **D** Interaction snapshots from simulation windows: Throughout the open state base stacking forms resulting in a more stable minima. Dotted orange lines highlight essential HB interactions.

base stacking with the template nt along with ”shaking” the motif F-R553/R555 interaction off triphosphate before transition toward the insertion configuration, in which RDV-TP can form very stabilized base pairing interactions with the template nt.

Additional close inspections on the RDV-TP local interactions with nearby residues show that the majority of HB and SB interactions are similar between the cases without and with forcing on the +1 template nt (Fig. 3.11C and Fig. 3.10C). Interestingly, one can identify that both HB and SB interactions from R555 and R553 (located on the motif F) with the triphosphate of RDV-TP, which are formed for the RDV-TP initial binding in the former stacking case with template forcing (see Fig. 3.10C), but become absent in the current case (without forcing on the template nt, Fig. 3.11C). Otherwise, the local HB/SB interactions with RDV-TP are highly similar for the two systems (Fig. 3.10 and Fig. 3.11), both initiated from the RDV-TP base stacking with the template nt binding configuration. Hence, in the RDV-TP insertion, the presence of the template forcing (or reduced fluctuations) along with the R555 (and R553) interaction with the triphosphate seems to hinder the RDV-TP insertion, which appears to be opposite to the trend in the ATP insertion (i.e., stronger R555/R553-ATP phosphate interaction in the open state under template forcing condition leads to a lowered ATP insertion barrier).

3.4 Discussion

The purpose of running the TMD simulations was to construct feasible dynamical paths of the nucleotide insertion to be used in the umbrella sampling simulations for the PMF construction, upon that the structural dynamics (with enhanced sampling in the transition state or barrier region) and energetics (or free energy profiles) of the insertion processes reveal and can be further compared. Our simulations first confirm that the nucleotide inserted or the closed form of the CoV-2 RdRp is indeed much more stabilized than the open form for nucleotide initial binding (about -3 to -5 kcal/mol), for RDV-TP or ATP. While the base pairing configurations of the initial binding ATP and RDV-TP are similarly stabilized (∼5 kcal/mol) relative to the corresponding closed insertion state, such an initial binding configuration is only dominant to ATP but not RDV-TP. Essentially, our calculations show that RDV-TP primarily forms base stacking with the +1 template nt rather than base pairing upon initial binding. Comparison between the RDV-TP insertion simulations conducted with varied initial binding configurations (stacking and base pairing) shows that motif A-D623 may stabilize the RDV-TP base stacking over the base pairing in the open state, by forming HBs with the sugar; in addition, motif C-S759 specifically forms HB with the RDV-TP sugar; S682 (near motif B) forms highly notable HB contact with the RDV-TP base, only in the base stacking configuration; motif F-K551 and K798 near the C-terminal of motif D stabilize the base stacking configuration by forming HB (or SB) interactions with the RDV-TP triphosphate,(also happen for ATP initial binding); motif G-K500 also forms HB with the +1 template backbone. Interestingly, as motif C-S759 does not form HB to ATP sugar upon initial binding, it forms HBs with the 3'-end of the primer RNA nt sugar in that case (see Fig. 3.12). Such interactions persist into the insertion states of both ATP and RDV-TP system. Additionally, in the case of RDV-TP base stacking with the template +1 nt, motif F-R555 also forms HB with the 3'-end of the RNA nt sugar (see Fig. 3.12). Consequently, the 3'-end of primer RNA nt cannot be involved with base stacking with RDV-TP, while R555 interaction with the phosphate of RDV-TP may also be prevented (see analyses later). The overall stabilization leads to ∼-2 kcal/mol (or ∼-3 kBT) relative initial binding free energy between the RDV-TP stacking and base pairing configuration. A docking stabilization energetics (∼-0.6 kcal/mol) between the RDV-TP and ATP was reported to a homology modeled CoV-2 RdRp,[64] and a similar energetic score revealed from our own docking trials (using the open form RdRp complex with RNA strands, see 2.4). Hence, it seems that the initial binding of RDV-TP to the CoV-2 RdRp can be about -2 to -3 kcal/mol more stabilized than ATP. An alchemical MD simulation for relative binding free energy calculation have presented a comparable stabilization energetics between RDV-TP and ATP (∼-2.8 kcal/mol) upon binding to the RdRp active site.[65] Nevertheless, the alchemical calculation was conducted in the absence of RNA, so it is unable to be compared in regard to the template RNA configuration. The computational results so far consistently point out that RDV-TP can bind to the CoV-2 RdRp active site (in an open form) more favorably than the natural nucleotide substrate ATP.

Nevertheless, the initial binding to RdRp only provides an initial nucleotide association and selection checkpoint to the nucleotide addition cycle or NAC. The followed insertion of the nucleotide to the active site becomes a next and likely the most important checkpoint in the NAC, in particular, for the single-subunit handlike RNA or DNA polymerases (RNAPs or DNAPs). In several such polymerase species, the nucleotide insertion is rate-limiting (or partially rate-limiting),[93;105;106] thus being critical for nucleotide selection.[94] Comparing to phage T7 RNAP we studied previously,[95;96] for which a substantial fingers

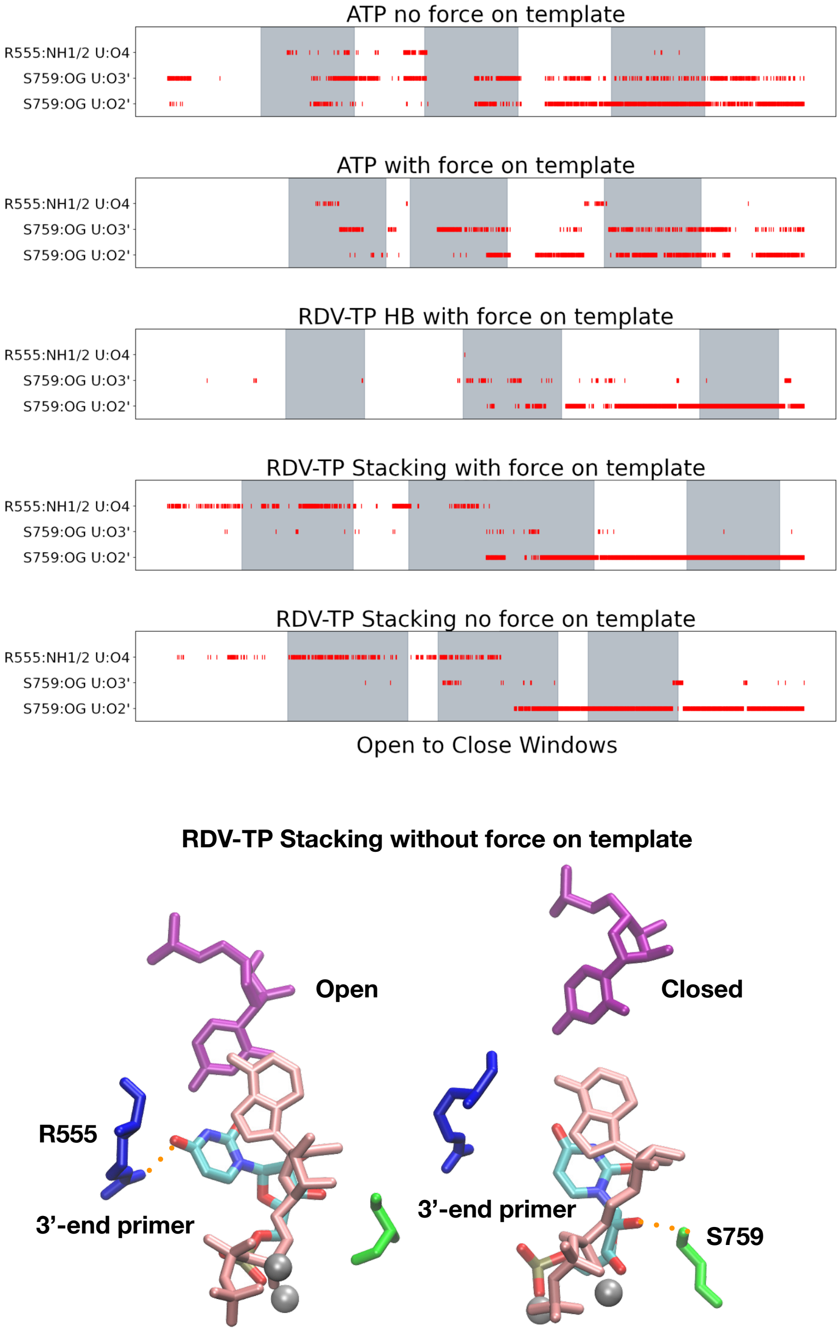


Figure 3.12: Hydrogen bond analyses around the 3'-end RNA primer nt (uracil base). (Upper) S759 forms frequent HBs with the 3'-end primer into both RDV and ATP insertion states, but only does that for ATP initial binding but not RDV-TP initial binding systems. R555 forms stable HB with 3'-end primer O4 oxygen in the RDV stacking initial binding state but no stable HB for RDV base pairing. (Lower) Open and closed states for RDV-TP from stacking insertion (no force on the template) showing the R555 and S759 HBs on the 3' end primer, respectively.

subdomain rotation happens with respect to the palm subdomain (from open to closed) during the nucleotide insertion, the viral RdRp conformational changes in accompany with the nucleotide insertion are mainly the active site distortions (from open to closed),[16] though remote residues on the structural motifs (A-G) can be more or less involved in the process. From the TMD simulations enforcing the CoV-2 RdRp from open to closed (see 3.1), we found that the motif A and D close similarly as that in PV RdRp.[13] Interestingly, the inserting ATP or RDV-TP has the base easily re-positioned toward the closed configuration in the TMD simulation, but has the triphosphate moiety hardly reaching to the targeted closed configuration. Hence, re-positioning of the triphosphate during the nucleotide insertion appears to link to events of crossing the free energy barrier. In current umbrella sampling simulations, the ATP or RDV-TP insertion barrier indeed depends on the relative template nt configuration or fluctuations, as well as local residue interactions with the triphosphate. In the ATP insertion, a comparatively low energetic barrier (∼2.6 kcal/mol) shows when the template nt is enforced or constrained to maintain stabilized base paring with ATP as if in the long-time unperturbed nucleotide insertion; the motif F-R555 interaction with the ATP phosphates along with motif A-D623 interaction with the sugar at the open state seems to facilitate the further ATP insertion. In comparison, for RDV-TP, the insertion barrier can be even lower (∼1.5 kcal/mol) when it is inserted without enforcing the template nt, so that the initial base stacking between RDV-TP and template nt can proceed freely to easily transit to the base pairing configuration into the closed insertion state. Contrary to the ATP insertion, motif F-R555/R553 close interactions (hydrogen bonding and salt-bridge) with the RDV-TP triphosphate in the open state appears to impede the RDV-TP insertion, which happens as the template nt is enforced in the simulation, no matter which initial configuration RDV-TP starts with the template (base pairing or stacking). Current simulations comparing RDV-TP and ATP thus suggest that the nucleotide insertion is coordinated by +1 template nt as well as the 3'-end of primer RNA nt with some notable interactions on the nucleotide upon initial binding, in which the triphosphate stabilization and re-positioning appear to be essential. It should be pointed out that the triphosphate reorientation of the incoming nucleotide had been suggested for the PV RdRp fidelity control.[107] Additionally, it is interesting to notice that motif F-R555 structurally corresponds to R174 from PV RdRp and R158 in HCV RdRp,[62] as well as to Y639 from T7 RNAP that is key to nucleotide selectivity and polymerase translocation.[95] Overall, the ATP insertion seems to be facilitated by an insertion path with quenched fluctuations on the +1 template nt for stabilized base pairing, while the RDV-TP insertion dominated by the template base-stacking populations, is supported by freely fluctuating template nt, leading to transition to the highly stabilized base pairing configurations, with an insertion free energy barrier as low as $\sim$1.5 kcal/mol or $\sim$2-3k$_B$T, marginally above thermal fluctuations.

Both the inserted ATP and RDV-TP can be then further stabilized well in the active site by the base paring interaction with the template nt. Though we haven't yet conducted energetic calculations to evaluate the relative stability between the RDV-TP and ATP in the insertion state, the equilibrium simulations of the insertion complexes of the two species suggest that the RDV-TP can be similarly or even more stabilized than ATP in the closed insertion state. There are also specific interactions that can well distinguish the natural nucleotide substrate from the nucleotide analogue in the insertion state: motif A-D623 forms specific HB contact with the ATP sugar but not with the inserted RDV-TP; K798 near motif D also closely interacts with the ATP gamma-phosphate into the insertion state, but not closely with that of RDV-TP; in contrast, motif B-T687 specifically forms HB with the RDV-TP sugar but not with that of ATP. The overall results thus suggest that binding/insertion of RDV-TP can be more facilitated than the natural substrate ATP to the active site of SARS-CoV-2 RdRp, seemingly consistent with in vitro measurements of the Michaelis-Menten constant K$_m$ obtained smaller for RDV-TP than for ATP, respectively.[58;59] If the nucleotide insertion is a single rate-limiting step (i.e., as in T7 RNAP[93]), then V$_{max}$ should also be significantly larger for RDV-TP than that for ATP, due to a lowest insertion barrier of the RDV-TP. However, the *in vitro* measurements of V$_{max}$ are similar for RDV-TP and

ATP.[58] Hence, other rate-limiting steps than the pre-chemical NTP insertion can exist in the NAC of the CoV-2 RdRp, e.g., the chemical catalysis, which may happen a bit slower for RDV-TP than ATP, so that overall the maximum elongation rates become similar for the two nucleotide species. More close examinations of stepwise kinetics of SARS-CoV-2 RdRp are therefore expected, ideally for both cognate and non-cognate nucleotide species, so that substantial information on the complete NAC as well as nucleotide selectivity could reveal. Note that following a successful RDV-TP incorporation to the end of viral RNA chain, additional nucleotide insertion still appears viable until the addition of the nucleotide downstream +3 to the incorporated RDV analog. Such mechanism has been suggested as a delayed chain termination of the nucleotide analogue,[58] which arises likely due to aberrant impacts of incorporated analogue on the synthesizing RNA chain in association with the viral RdRp, together with failure of ExoN cleavage or proofreading to the nucleotide analogue.

3.5 Conclusions

Via modeling and all-atom MD simulation, we found that remdesivir nucleotide analogue can bind to the open active site of SARS-CoV-2 RdRp via base stacking with the +1 template nt. Such a stacking configuration appears to be more stabilized, relative to the insertion state, than the Watson-Crick base pairing configuration formed between ATP and the template uracil base. Umbrella sampling simulations further show that the remdesivir analogue stacking with the fluctuating template then inserts or transits to form high-stabilized base pairing with the template as the active site closes. The corresponding insertion barrier for remdesivir analogue can be even lower than that of a low-energetic path of the ATP insertion, during which the template forms stabilized base pairing with ATP. Additionally, our analyses on hydrogen bonding and salt-bridge interactions during the nucleotide or analogue insertion show that (i) the initial remdesivir base stacking can be particularly stabilized by motif A-D623 along with motif C-S759 with sugar, S682 with base, and motif G-K500 with the template, motif F-K551 and K798 with phosphate, as well as motif F-R555 with 3'-end primer; (ii) insertion of remdesivir analogue can be facilitated by thermal fluctuations but hindered by motif F-R555/R553 interaction with the triphosphate, while insertion of ATP is made easier by lowering fluctuations and taking advantage of the R555/R553 interaction with the triphosphate; (iii) the inserted remdesivir analogue and ATP are distinguished by specific sugar interaction via motif B-T687 and motif-A D623, respectively. Such findings also reveal potential SARS-CoV-2 RdRp fidelity control via particular residue interactions with the nucleotide substrate sugar, base, and triphosphate moieties, along with +1 template coordination.

Chapter 4

dATP and GTP Vs ATP and RDV: A Comparative Study of Initial Binding and Insertion Through Equilibrium Ensemble Simulations

This chapter focuses on the initial binding and insertion states of noncognate dATP and GTP, with the objective of elucidating the mechanisms by which CoV-2 RdRp manages the binding of noncognate NTPs. We continue to utilize the models developed in previous chapters, as, at the beginning of this study, no structures of insertion or initial binding states were available. Our investigation begins with the generation of a large equilibrium ensemble for the noncognate NTPs. Measurements derived from this ensemble are then compared with those from our previous studies on ATP and RDV-TP, which also utilized an equilibrium ensemble. In total, we generated 8 μs of simulation data, enabling a systematic exploration of the CoV-2 RdRp. The insights obtained from this work reveal distinctive behavior patterns of noncognate NTPs, particularly in the initial binding state, which later proved instrumental in our free energy calculations.

4.1 Introduction

The SARS-CoV-2 virus responsible for the COVID-19 pandemic continues to evolve[108] and pose a threat to human life.[109] While the vaccines developed are demonstrating success, there remains an imperative to accelerate antiviral development on therapeutics, considering virus's evolution and resistance, vaccine hesitancy among individuals, or the inability of some countries to afford the available vaccines. While much of the drug development is focused on targeting the viral spike protein[110–112] or the main protease,[113–115] there are significant challenges. The spike protein is known for its high variability[108;116;117] and the protease[118] can also mutate, becoming resistant to drugs, as seen in HCV.[119;120] In contrast, the core replication machinery of SARS-CoV-2, the RNA dependent RNA polymerase (RdRp) or nonstructural protein nsp12, is a highly conserved drug target.[121;122] Here we focus on studying SARS-CoV-2 RdRp to understand its underlying function and mechanism that is crucial for future drug development,[123] particularly given the significant gaps in knowledge regarding the functioning of viral RdRps.[124;125]

Upon the pandemic upheaval in 2020, a few high-resolution cryo-EM structures of SARS-CoV-2 RdRp were released immediately, including apo forms with the RdRp active site open[7;11] and a post-catalysis form with the active site closed.[11] The post-catalysis structure was bound with a nucleotide drug analogue from remdesivir (RDV).[11] These structures were in complex with segments of the nsp7/nsp8 cofactors. Later, additional high-resolution structures were resolved with longer nsp8 being "sliding poles",[68] as well as RdRp in conjunction with the nsp13 helicase enzyme,[87] and in both pre- and post- translocation states.[8] Further structures illustrated RdRp backtracking[88] or stalling[126] due to the interaction with the drug analogue RDV. Similarly, there were structures obtained with favipiravir, another nucleotide analogue drug.[127;128] Overall, these structures adopt the post-catalysis state[11] in the nucleotide addition cycle (NAC) of the viral RdRp, leaving the initial nucleotide binding (active site open) and pre-catalytic insertion or substrate (closed) states unresolved. It was until very recently that the insertion state with ATP as a cognate nucleotide triphosphate (NTP) bound in the active site was resolved.[10] Currently, the initial binding or open state structure of RdRp remains unresolved, although such a structure has been identified for the Poliovirus[16] (PV) or Enterovirus[24] (EV) RdRp, which share structural similarities with SARS-CoV-2 RdRp.

Efforts to identify drug inhibitors for the SARS-CoV-2 RdRp have been made extensively.[52;55;123] Computational docking has often been employed, which predominantly focus on nucleotide immediate binding to an apo-form RdRp structure,[129;130] non-differentiating the active open or closed form, and overlooking initially the functional RdRp (nsp12) elongation complex composed additionally of nsp8, nsp7, and RNA strands. Moreover, while atomic molecular dynamics (MD) studies have provided insights into interactions of nucleotide analogues with the viral RdRp, they often utilize directly the insertion state,[131–133] ignoring the initial nucleotide analogue binding stage that can be essential for nucleotide screening or selectivity upon entry. Meanwhile, single-molecule studies have offered a glimpse into the dynamics of the elongation cycle, revealing that RdRp can adopt fast, slow, or very slow catalytic pathways with variable rates contingent upon the kinetic pathway.[134] Additionally, information in regard to the RdRp translocation in the NAC has advanced through the cryo-EM studies, which demonstrated a structural rearrangement in nsp8 to accommodate the exiting RNA duplex.[8] Computational work has shown that the incorporation of RDV-TP into the primer strand results in a steric clash at the conserved motif-B of the RdRp leading to an unstable post-translocation state in comparison with pre-translocation, i.e., as a mechanism for antiviral analogue termination of elongation.[135] An alternative suggestion based on single-molecule studies[134] proposes that RdRp backtracks up to 30 nucleotides (nts) after RDV-TP incorporation, which can be interpreted as elongation termination in standard assays. Despite these efforts on quantitative studies of the RdRp NAC, a critical gap remains in understanding initial nucleotide substrate binding to insertion, which is fundamental for nucleotide selectivity and antiviral drug design, given the substrate screening and pre-chemistry inhibition as essential fidelity checkpoints in stepwise NAC.[22;23;25] We aim to further unravel the intricacies of RdRp's nucleotide selectivity by modeling the open and closed states for noncognate NTPs: GTP and dATP. It's worth noting that GTP can indeed form stable base pairing with a uracil mismatch through wobble base pairing, which has thermodynamic stability comparable to an A·U Watson-Crick base pairing.[136] In theory, dATP, possessing the appropriate adenine base, should form Watson-Crick base pairing. However, we anticipate its rejection due to the absence of a 2'-OH group on the ribose sugar.

4.2 Computational Details

4.2.1 Construction of Initial Binding and Insertion States of Noncognate: dATP and GTP

Based on the approach used in previous work on modeling SARS-CoV-2 RdRp,[50] we constructed models for initial binding (open) and insertion (closed) of various NTPs using high-resolution Cryo-EM structures of SARS-CoV-2 nsp12-nsp7-ns8-RNA complexes (see 2.2.1). The model for RDV-TP in the insertion state was created using a post-catalytic structure bound with RDV-TP (PDBid:7BV2) as a reference, which also includes catalytic Mg^{2+} ions.[11] For the ATP insertion model, we positioned ATP over the RDV-TP in the insertion model. The RDV-TP initial binding model was created using an apo structure (PDBid:7BTF)[7] aligned with the RDV-TP insertion structure, with RNA, RDV-TP, and Mg^{2+} ions copied over. The ATP initial binding model was built following the same process as described above. The dATP states were built using the ATP models as reference and removing the 2' OH group. Alternatively, GTP was aligned with the ATP models with subsequent geometry optimization to force wobble (WB) base pairing (see **Figure 4.3** *lower right*).[136] For complete details on preparation of RdRp model in association with RDV-TP (with force-field parameterization) and ATP, please see previous work.[50;66]

4.2.2 Simulation Parameters and Setup

Simulation parameters follow the same setup described in section 2 simulation details 2.2.4. All MD simulations were performed using GPU-accelerated Gromacs 2021 software[79] with the following forcefields: Amber14sb,[35] Parmbsc1,[40] triphosphate parameters developed previously.[80] Enhanced or umbrella sampling methods were performed using a Gromacs 2021 package patched with PLUMED.[137] Each complex was solvated with explicit TIP3P water[38] in a cubic box with a minimum distance from complex to the wall of 15 Å. Resulting in an average box dimensions of 15.7nm x 15.7 nm x 15.7 nm. The overall negative charge of the complex was neutralized and ions were added to create a salt concentration environment of 100mM. Full simulation systems were ∼382,000 atoms in size. A cut-off of 10Å was used to treat short range electrostatics interactions and the Particle-mesh-Ewald (PME) algorithm to treat long range interactions.[82] The LINCS algorithm is used to constrain bonds to hydrogen atoms allowing the use of a 2 fs timestep when integrating the equations of motion.[79] Temperature was kept at 310K using the velocity rescaling thermostat. Pressure was kept at 1 bar using Berendsen barostat[84] during equilibration and Parrinello-Rahman barostat[138] for production, targeted MD (TMD), and umbrella simulation runs. Each system was minimized for a maximum of 50000 steps using steepest-descent algorithm, followed by a slow equilibration with restraints released (every 1 ns) going from NVT (canonical or constant volume and temperature) ensemble to NTP (constant pressure and temperature) as previously used.[50]

For each NTP initial binding and inserted states ten 100 ns equilibration trajectories were launched, with 10 ns removed from the start, to create 1 μs for each NTP state, totaling ∼8 μs of simulation time for RDV-TP/ATP/dATP/GTP in both open and closed conformation states. The generated equilibrium ensemble was then used for analysis and selection of references for free energy calculations.

4.3 Analysis

4.3.1 RMSD

The RMSD is measured for each RdRp complex on the subdomains (protein backbone), RNA (phosphate backbone), and NTP (heavy atoms) as on well as the the key motifs (from A to G) interacting with each NTP Fig. 4.1, Fig. 4.2. Both initial binding and insertion equilibrium ensemble trajectories were aligned via the finger's subdomain to their respective initial states after minimization for each NTP.

4.3.2 Base Pair Geometry

Base Pair Geometry was measured between the incoming NTP and uracil template nt at +1 by calculating the base plane angle (C1'-C7-C5 for NTP and C1'-C2-C5 for template) and the distance between the center of mass (COM) of each base (Fig. 4.3). A similar protocol was followed for measuring the geometry between each NTP and the 3' end primer: measuring the COM between bases and the base plane angle for (C1'-C7-C5 for NTP and C1'-C2-C5 for 3' end primer) (Fig. 4.4). Measurements were conducted using the MDAnalysis python package[103] and gromacs gangle module.

4.3.3 Hydrogen Bond Occupancy

Hydrogen bonds are measured using the gromacs 2021 module. A distance cutoff of $\leq$3.5 $\mathring{A}$ between donor-acceptor involved and hydrogen-donor-acceptor angle cutoff of $\leq$ 30° is used as the criterion to indicate a proper hydrogen bond. Unique hydrogen bonds with an occupancy greater than 20% (within a 900 ns combined trajectory) are considered for analysis.

Plots are created using python packages seaborn[139] or matplotlib.[140]

4.4 Equilibrium ensemble simulations: protein structural variations and distinctive dynamical responses to different NTPs upon initial binding and insertion

Upon modeling the initial binding (open) and insertion (closed) complexes for the four NTPs, we conducted equilibrium all-atom MD simulations of 10 x 100 ns for each system. We began by measuring and comparing the RMSDs of the protein subdomains, RNA and NTP's using the respective energy minimized structures for insertion or initial binding as references (see Fig. 4.1). For all four NTP binding, the fingers subdomain RMSD is centered around ~1.3 Å and the palm subdomain remains similarly aligned with fingers, especially in

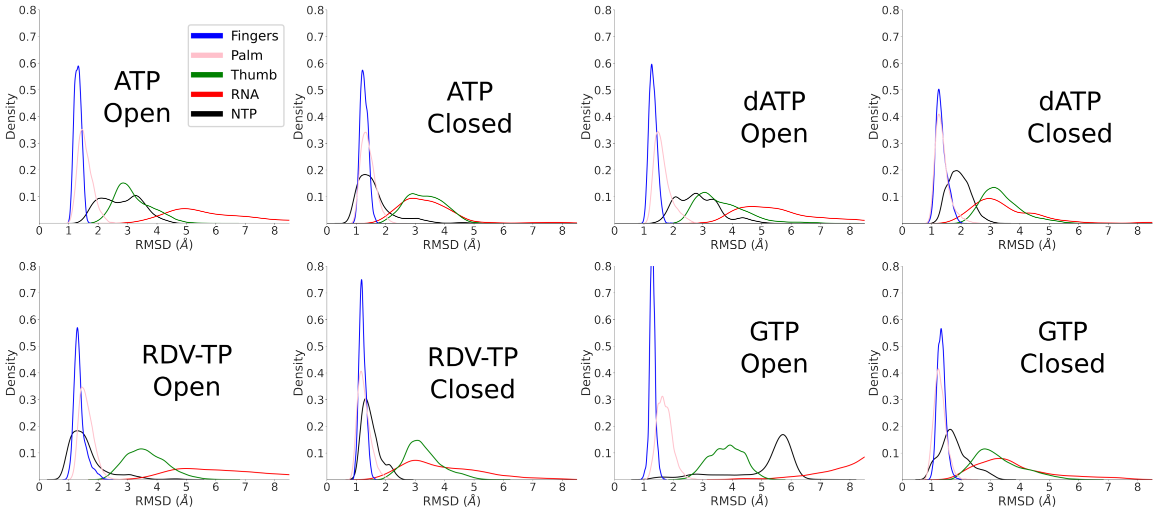


Figure 4.1: The root-mean-square displacements (RMSDs) of RdRp structural subdomains (backbone atoms), RNA (phosphate backbone), and NTP (heavy atoms) measured from ensemble MD simulations. The RMSDs are shown for ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) upon initial nucleotide binding (active site open; left) and insertion (active site closed; right) states. The subdomains are shown in different colors: fingers (blue). Palm (pink), thumb (green), RNA (red), and NTP (black).

the insertion state. The thumb subdomain has a comparatively wide distribution of RMSD in all cases, indicating conformational flexibility. All NTP display less variability in the insertion state, along with the fingers/palm subdomain. The cognate ATP and RDV-TP analog both exhibit lower RMSDs than non-cognate dATP or GTP in the insertion state. For initial binding, the RMSDs of GTP along with the RNA scaffold are much larger than that for the other NTP binding, showing significant dynamical responses to the mismatched GTP. Additionally we compared the RMSDs of the seven conserved motifs (see Fig. 4.2 and Table. 4.1). Both motif B and C show comparatively low RMSDs in the insertion state. Motif-C demonstrates notable stability among all motifs for all inserted NTPs, likely due to it hosting the key catalytic residues (S759/D760/D761). The RMSDs of other structural motifs displayed significant variations upon initial binding of different NTPs. In the insertion state, motifs F and G show more distinctions dynamically than rest of motifs for different NTP substrates.

Next, we examined the base association or pairing geometry between the initially bind-

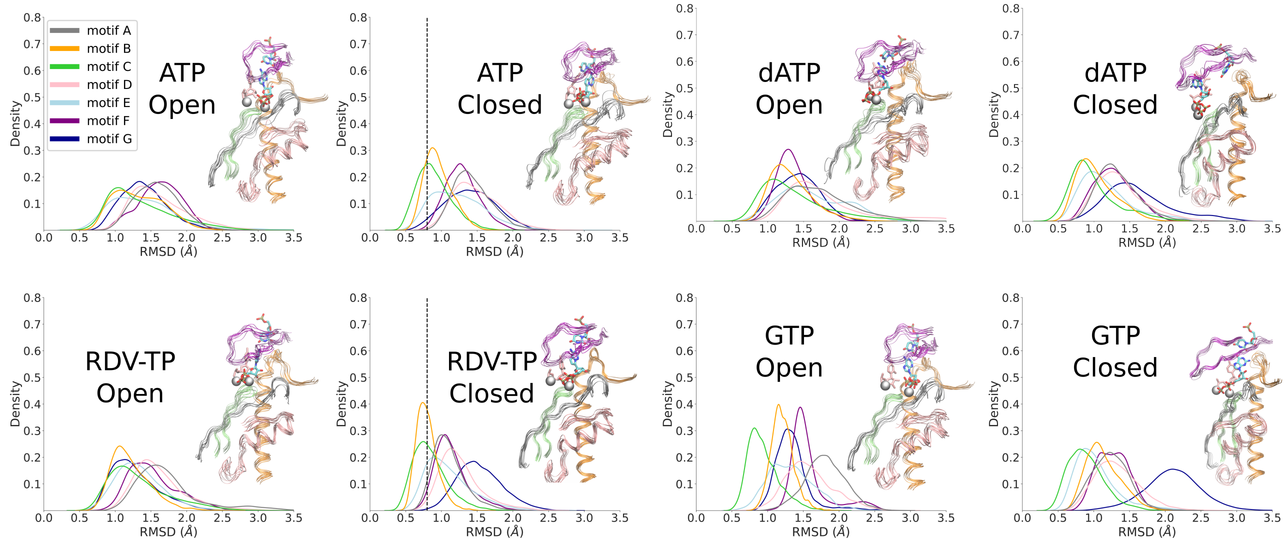


Figure 4.2: The root-mean-square displacements (RMSDs) of RdRp structural motifs (backbone atoms) measured from ensemble MD simulations. The motif RMSDs are shown for ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) upon initial nucleotide binding (active site open; left) and insertion (active site closed; right) states. The seven key motif RMSD are displayed in different colors: motif A (gray), Motif B (orange), Motif C (green), Motif D (Pink), motif E (light blue), motif F (purple), and motif G (dark blue). Structural representations of those motifs along with NTP, uracil template nucleotide, and two catalytic MG ions are shown for each simulation system. The dotted black line indicates the reference group of motifs (B & C) for systems of inserted ATP/RDV-TP.

ing/insertion NTP and template +1 nt (Fig. 4.3). In the insertion equilibrium ensemble (active site closed), we observed that the NTP-template distance distribution predominately centers at ∼6.5 $\AA$ and base plane angle ∼30°, resulting in either the stable Watson-Crick (WC) base pairing (for ATP/RDV-TP/dATP-template) or wobble base pairing (WB) interactions (for GTP-template uracil). The probability of WC/WB base pairing is high for ATP (69%), RDV-TP (94%) and GTP (70%), low for dATP (47%). Indeed, dATP upon insertion displays more flexibility than other NTPs in association with the template nt (see Fig. 4.3 *upper right*). In the NTP initial binding ensemble (active site open), the NTP-template association geometries vary significantly. Upon ATP initial binding, as a significant amount of WC population (48%) is identified, a comparable amount of un-paired or weakly paired

RMSD	Initial Binding "Open" State				Insertion "Closed" State			
Motifs	ATP	RDV-TP	GTP	dATP	ATP	RDV-TP	GTP	dATP
A	1.6±0.3	1.7±0.5	1.8±0.3	1.7±0.4	1.4±0.3	1.1±0.2	1.1±0.2	1.3±0.3
B	1.3±0.4	1.2±0.3	1.2±0.2	1.3±0.3	1.0±0.2	0.8±0.2	1.1±0.2	1.0±0.3
C	1.4±0.5	1.3±0.4	0.9±+0.2	1.4±0.5	0.9±0.2	0.8±0.2	0.8±0.2	1.0±0.3
D	1.6±0.4	1.6±0.4	1.6±0.3	1.8±0.6	1.5±0.3	1.3±0.3	1.4±0.3	1.4±0.4
E	1.4±0.5	1.3±0.3	1.3±0.3	1.6±0.5	1.2±0.4	1.1±0.3	1.0±0.3	1.2±0.4
F	1.6±0.3	1.5±0.3	1.6±0.3	1.4±0.2	1.3±0.2	1.1±0.2	1.3±0.2	1.3±0.3
G	1.4±0.3	1.3±0.3	1.3±0.2	1.4±0.3	1.4±0.3	1.5±0.3	2.1±0.4	1.6±0.5

Table 4.1: Average Motif RMSD from the initial binding and insertion equilibrium ensemble simulations using the closed state minimized structure as reference. Units are in Angstrom.

(single HB) ATP-template uracil configurations are also present. Upon initial binding of RDV-TP, three configurations have been identified, with either the WC base pairing (36%) or base stacking (38%) being stabilized, and (26%) unpaired.[50] For non-cognate NTPs, the initially bound dATP marginally forms WC base pairing (12%) with the template nt, while GTP upon initial binding cannot forms stabilized WB base pairing with the template uracil.

Additionally, we measured the geometry between the NTP and 3' end primer (Fig. 4.4). From different NTP insertion ensembles, the NTP-3' end primer distribution centers closely around ∼4-6 $\AA$ and a base plane angle ∼150° to 170°, showing stability. In contrast, upon initial binding, the mismatch GTP associated with the 3' end primer in largely diverse geometries (distance spans from 4 to 13$\AA$ and the angle varies from 20° to 170°). The noncognate dATP upon initial binding also shows diverse geometries in association with the 3' end primer (distance spans from 4 to 9 $\AA$ and the angle varies from 40° to 180°). For cognate ATP and RDV-TP, the initial binding geometries with respect to the 3' end primer are much more localized, i.e., the distance dominantly spans from 4 to 5 $\AA$, with small populations in between 8-10 $\AA$; the angle varies from around 100° to 180° below than the non-cognate species. In summary, in the active site closed state or insertion equilibrium ensemble, various NTPs display much less variation of geometries with respect to the template +1 nt or the 3' end primer than those in the active site open state or initial binding equilibrium ensemble. The highly diverse configurations of initially bound NTP (particularly non-cognate NTP) suggest that detection of various species of incoming NTP starts well from the beginning, i.e., upon the initial NTP binding when the active site remains open.

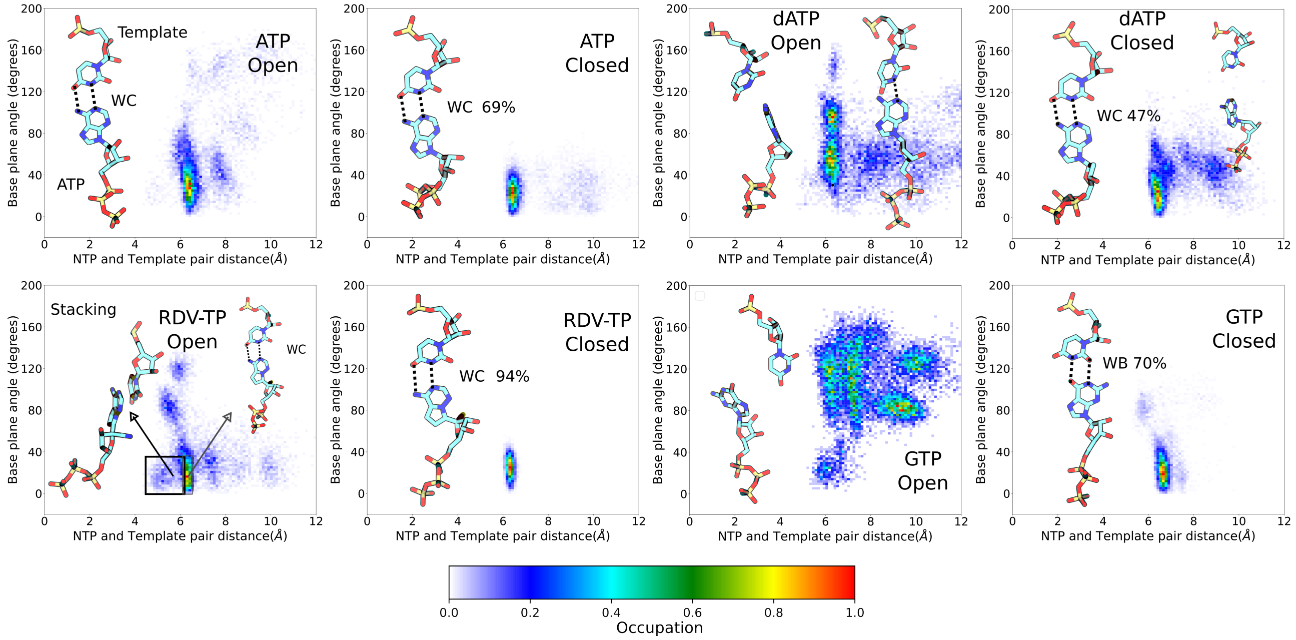


Figure 4.3: NTP-template association geometries from ensemble equilibrium simulations. The geometric measures (see **Methods**) between the uracil template +1 nucleotide (nt) and individual incoming NTP are displayed, for ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right), upon initial binding (left) and insertion (right) for each NTP species. Licorice representations of the NTP and template nt show the dominant geometries for each simulation system. Dotted lines indicate the hydrogen bonds for the standard Watson-Crick (WC) or wobble base (WB) pairing.

Furthermore, we also measured the HB occupancies for various NTPs in respective associations with the protein, template +1 nt, and 3' end RNA primer (histogram statistics shown in Fig. 4.5), for both the NTP initial binding and insertion equilibrium ensembles. We observed an increase in HB occupancies between the NTP triphosphate tail and protein from initial binding to the inserted ensembles for ATP and dATP, with more protein-sugar HBs for the inserted ATP (with D623 and N691) than for the inserted dATP. On the other hand, the inserted RDV-TP ends up with fewer protein-triphosphate HBs than the inserted ATP but much stronger HBs (with T687 and N691) in the protein-sugar association. Meanwhile, GTP insertion led to many but weak (low occupancy) HBs formed, indicative of some

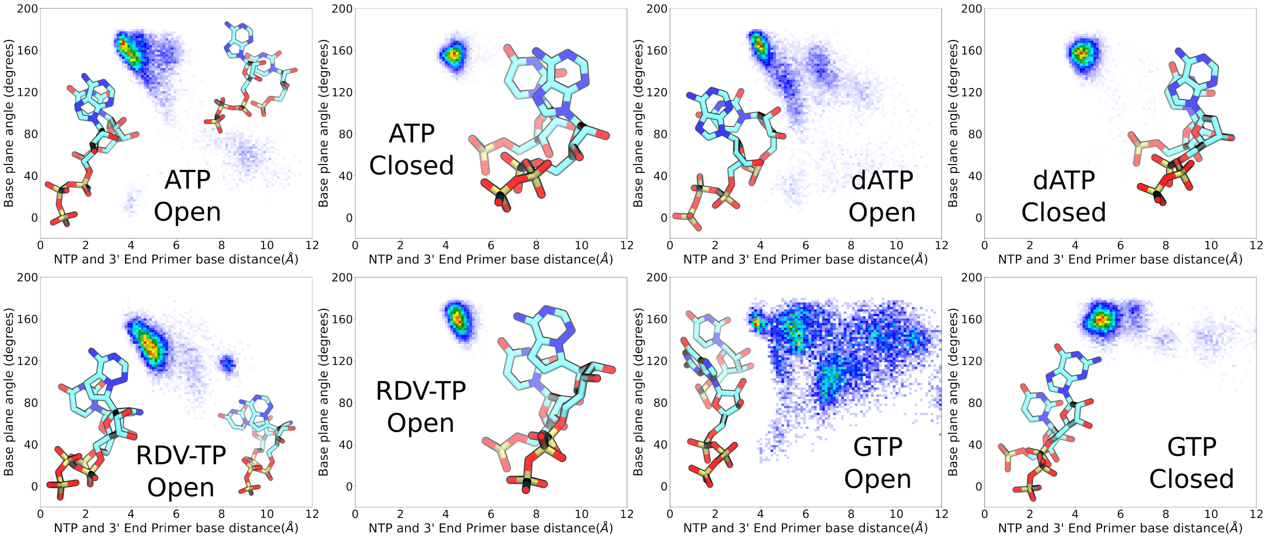


Figure 4.4: NTP and 3'-end primer association geometries sampled from equilibrium ensemble simulations. The geometric measures (see Methods) between the 3'-end primer and individual incoming NTP: ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) are demonstrated, upon initial binding (left) and insertion (right) for each NTP species. Licorice representations of the NTP and 3'-end primer show the dominant geometries for each simulation system. Distance is measured by the center of mass between the bases. Base plane angle is measured using the C1'-C2-C5 (3' end primer) and C1'-C7-C5 (NTP).

instability.

Additionally (see Fig. 4.6), in the open state, several protein residues (motif-F K551, K545, A558 and motif-B S682) form HBs with template +1 nt upon initial binding of GTP. While protein S501 (from motif G) forms HB with template +1 nt strongly in the presence of ATP (71%) and RDV-TP (78%), or marginally upon dATP (40%) and not present in GTP. In the closed state, the S501-template HB persists and becomes highly stabilized for every NTP insertion state. These findings again reflect distinctive HB patterns formed around the active site upon association of various NTPs, from initial binding to insertion. Nevertheless, due to various populations of NTP binding configurations, especially in the non-cognate initial binding state, it is not clear which HB interactions contribute to stabilize or destabilize cognate vs ncNTPs for nucleotide selectivity.

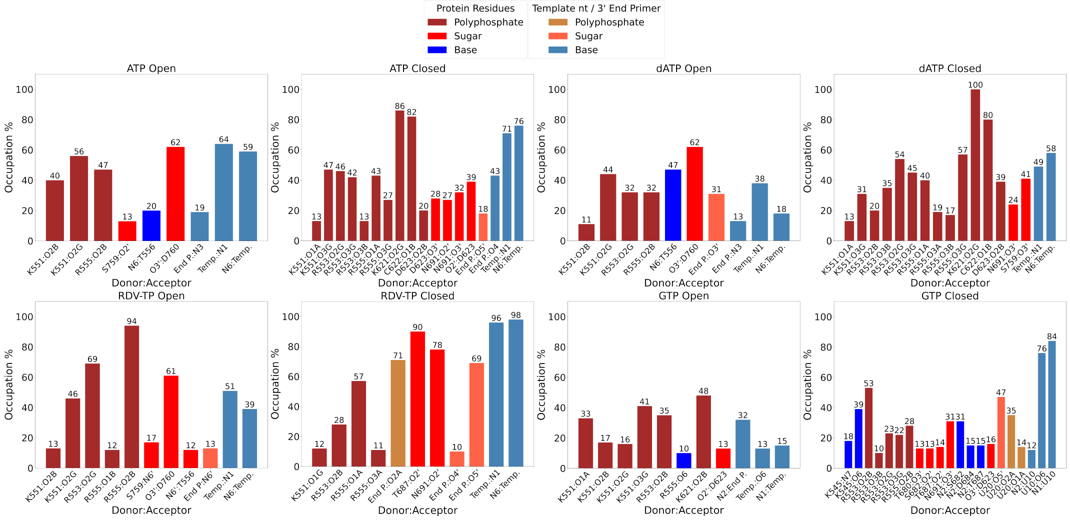


Figure 4.5: The hydrogen bond occupancy of the equilibrium ensembles for each NTP: ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) are demonstrated. Each unique interaction >10% population is considered. Two color-code sets are used: protein-NTP interactions use brown (polyphosphate), red (sugar), and blue (base), and for template-nt / 3' end primer-NTP interactions use light brown (polyphosphate), light red (sugar), and light blue (base).

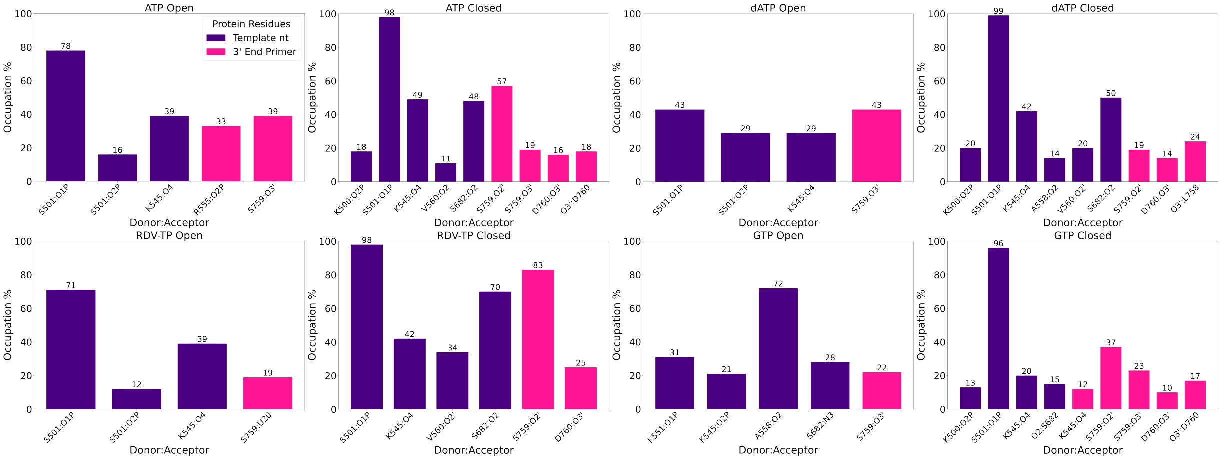


Figure 4.6: The hydrogen bond occupancy of the equilibrium ensembles for protein-template nt Uracil (purple) and protein-3' end primer (pink) for each NTP simulated: ATP (upper left), RDV-TP (lower left), dATP (upper right), and GTP (lower right) are demonstrated, upon initial binding (left) and insertion (right) for each NTP species. Each unique interaction >10% population is considered.

Chapter 5

Nucleotide Selectivity: Comparative Assessment of Insertion Free Energy for dATP, GTP, ATP, and RDV

In this chapter, we undertake a comparative analysis of the computed free energy of insertion between noncognates dATP and GTP. In addition, we can provide a more comprehensive interpretation of the mechanism and draw comparisons with ATP and RDV-TP. At this stage, we also conducted alchemical calculations, which facilitated the ordering or ranking of insertion potential of mean forces (PMFs). As discussed previously 3.3 the cognate ATP and drug analogue RDV-TP exhibit greater stability in the closed state, resulting in a negative ΔG. In contrast, the noncognate nucleotides show greater stability in the initial binding phase, yielding a generally positive ΔG.

5.1 Introduction

Accompanied with the nucleotide substrate binding to insertion, a substantial protein conformational transition happens, which is likely to be a rate limiting step in the NAC, as demonstrated in structurally similar single-subunit RNA or DNA polymerases (RNAPs or DNAPs).[141–143] Such a rate-limiting pre-chemical step accordingly plays a highly essential role in the nucleotide substrate selectivity.[94] To quantify the process with energetics, we calculated the free energy profile or the potential of mean force (PMF) of the nucleotide insertion, recently for cognate substrate ATP and the corresponding nucleotide drug analogue RDV-TP.[50] We found that both ATP and RDV-TP become significantly more stabilized in the insertion state than upon initial binding. In contrast to natural substrate ATP, RDV-TP behaves differently in its interaction with the template nt. Cognate ATP forms the Watson-Crick (WC) base pairing with template uracil at the +1 position, from the initial binding through to insertion. On the other hand, RDV-TP initially forms base stacking with the template nt at +1 upon binding. It then inserts into the active site, facing an energetic barrier that is marginally low or comparable to that of cognate ATP.[50]

Note that RDV-TP analogue differs from the cognate ATP by only a few atoms: with a 1' cyano group attached on the sugar C1' and 3 atomic replacements on the base. Interestingly, we have also noticed that one conserved motif-F (R553+R555) essentially facilitates the insertion of cognate ATP via interactions with the ATP-triphosphate tail. While such phosphate interactions could potentially hinder the drug analogue RDV-TP insertion, it was subtly avoided upon sufficient thermal fluctuations from the template nt+1 that forms base stacking with RDV-TP.[50] In current work, we focus on characterizing the intrinsic nucleotide polymerase enzyme selectivity of SARS-CoV-2 RdRp, i.e., the selectivity or differentiation between natural cognate NTP (ATP here) and natural non-cognate NTP substrates. To do that, we examined the nucleotide insertion dynamics of non-cognate dATP and GTP, in comparison with cognate ATP and RDV-TP analogue, and calculated the insertion PMFs of dATP and GTP starting from initial binding stage. Since the polymerase NAC lasts over tens of milliseconds in general,[134] the rate-limiting transition accompanies the nucleotide binding to insertion (or active site from open to closed) of RdRp is expected to be on the millisecond timescale.[141–143] Therefore, such an insertion process cannot be sampled directly using equilibrium MD simulation that is limited by the sub-microsecond to microseconds timescale.[44;144] To obtain free energetics of such a dynamics process, we extended our previous methodology on employing umbrella sampling MD simulations to construct the PMFs of various NTPs from initial binding to insertion.[50;96] We first constructed atomic structural models of the RdRp-nsp7-nsp8-RNA complexes bound with the noncognate nucleotide (ncNTP) species, in both initial binding and insertion states. Subsequently, we performed all-atom equilibrium simulations at sub-microseconds in ensembles to characterize the respective initial binding and insertion states of the non-cognate dATP/GTP bound RdRp complexes. Lastly, we obtained the PMFs of the dATP/GTP insertion processes using the umbrella sampling MD simulations, following initial insertion paths constructed on top of

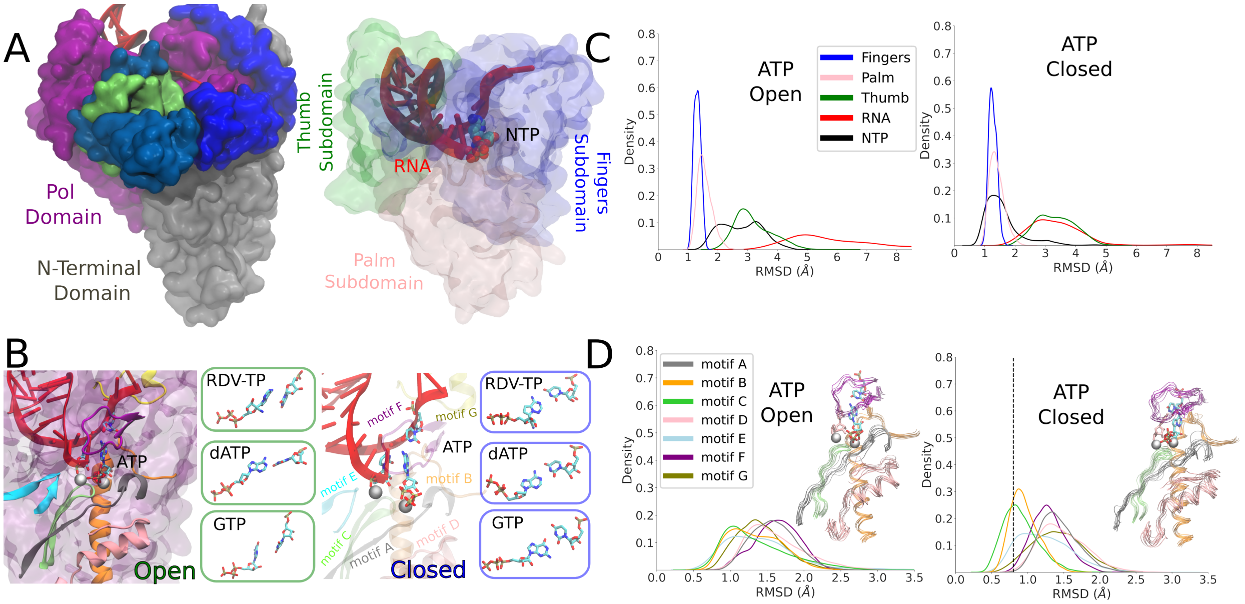


Figure 5.1: The constructed structural models of the SARS-CoV-2 RdRp elongation complex in its initial NTP substrate binding (open) and then inserted (closed) states. Top left: The simulated elongation complex is depicted with the Nsp12 pol domain (purple) + N-terminal domain (gray), two cofactor nsp8 (blue) and nsp7 (green), RNA (red) and an NTP in the active site. Top center: The three major subdomains within the pol domain fingers in blue, palm in pink, and thumb in green. The RNA is also shown in red as well as the NTP shown in space filling spheres. Top right: The subdomains RMSD for initial binding and insertion states for ATP highlighting the subtle fingers and palm coming together. See Fig. 4.1 for the rest of the NTP's. Bottom Left: The modeled and simulation equilibrated ATP is shown as bound initially to an open active site, with the seven protein motifs highlighted in color. The boxes to the right show the initially bound RDV-TP, dATP, and GTP that were also modeled and equilibrated from the simulations. Bottom Right: The modeled and equilibrated ATP is shown in the insertion or the active site closed state, with the seven structures motifs shown as well, and the boxes to the right displaying the modeled and equilibrated insertion configurations of RDV-TP, dATP, and GTP to the closed active site.

collective reaction coordinates (RCs), according to displacements of atomic coordinates from seven highly conserved structural motifs of RdRp (A to G) and incoming NTP (with and without the associating template nt). Our aim is to elucidate the intrinsic nucleotide selectivity of SARS-CoV-2 RdRp, which turns out to be primarily relying on 'trapping' the non-cognate nucleotide species upon entry or initial binding to certain configurations at the active site and preventing them from further insertion.

5.2 Reaction Coordinate for GTP and dATP

The reaction coordinate (or RC) used in Umbrella Sampling (U.S.)[45;49] is the difference in RMSD (eq. 5.2) with respect to two reference structures, one in the open (NTP initial binding) and the other in the closed (NTP insertion) state, respectively.

$$\begin{aligned} RC(X) = \delta RMSD(X) = & RMSD(X, X_{\text{Open ref}}) \qquad (5.1) \\ & - RMSD(X, X_{\text{Closed ref}}) \qquad (5.2) \end{aligned}$$

The reference structures (Open/Closed ref) were selected using a $\delta RMSD$ equation on the the first half (50 ns) of the equilibrium trajectory with respect to the initial models (see 3.2.1). However, in the case of non cognates we visually inspected trying to select a base pair geometry from the well sampled regions Fig. 4.3.

After selecting the reference structures, a path is generated for U.S. using target molecular dynamics (TMD)[98] simulations. We create a forward path (open to closed ref) and a backward path (closed to open), applying force along the following atomic coordinates (X): nsp12 motifs (motif A-G backbone atoms), NTP (heavy atoms), and finally as tested previously[50] two protocols: i) with force on the template nt +1 ii) without force on the template nt +1 (see 3.2.2).

From the TMD paths created between the two reference states, intermediate structures were evenly (every 0.1 Å along the RC) selected to be used in launching umbrella sampling simulations. The force constants used in the TMD were carried over for U.S. simulations (see Table. 5.1). The collected RC value histograms were then re-weighted using the weighted histogram analysis method[99] as implemented by the WHAM package.[101] For each set of trajectories making up the umbrella windows, 10-ns trajectory was removed from the start, followed by convergence check for every 10ns. For dATP and GTP, the convergence time ranged from 50ns-60ns (Fig. 5.2) for the case without force on the template. Additionally, error bars are estimated using the bootstrapping error analysis method[102] implemented in

Table 5.1: Umbrella Sampling Parameters: force constant k for each path and total number of windows used

NTP	Forward k $\left(\frac{\text{kcal}}{\text{mol}\mathring{A}^2}\right)$	Backward k $\left(\frac{\text{kcal}}{\text{mol}\mathring{A}^2}\right)$	# of Windows
GTP	501.9	501.9	24
GTP†	250.95	501.9	24
dATP	501.9	250.95	21
dATP†	501.9	250.95	35

† Denotes without force applied on template nt +1

the WHAM package. For a thorough description on utilizing the umbrella sampling method for constructing the PMFs for ATP and RDV-TP insertion see 1.3.2 and 3.2.2.

To additionally examine the insertion structures simulated from umbrella sampling, we used steered molecular dynamics (SMD) to check whether the stability of the inserted GTP/dATP show consistencies with the umbrella sampling or the PMFs constructed. The SMD was implemented controlling two center of mass (COM) distances defined as between the heavy atoms of the NTP and another COM with the C$_\alpha$ atoms from residues within 10 Å of the 3' end primer. The distance between the two COMs was increased at a slow rate of 1 Å per 100 ns, with a force constant set to 2.4 $\frac{kcal}{mol\mathring{A}^2}$.

5.3 Results

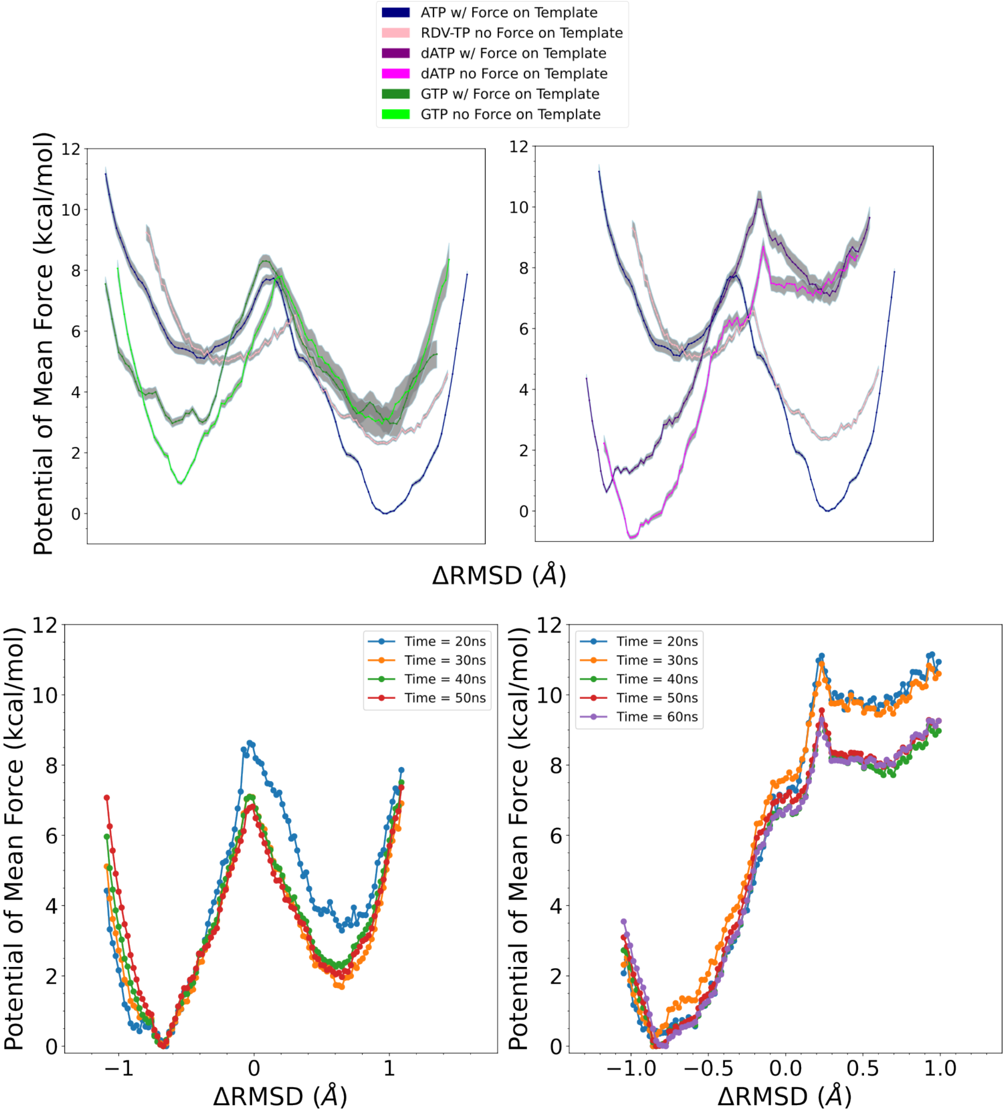


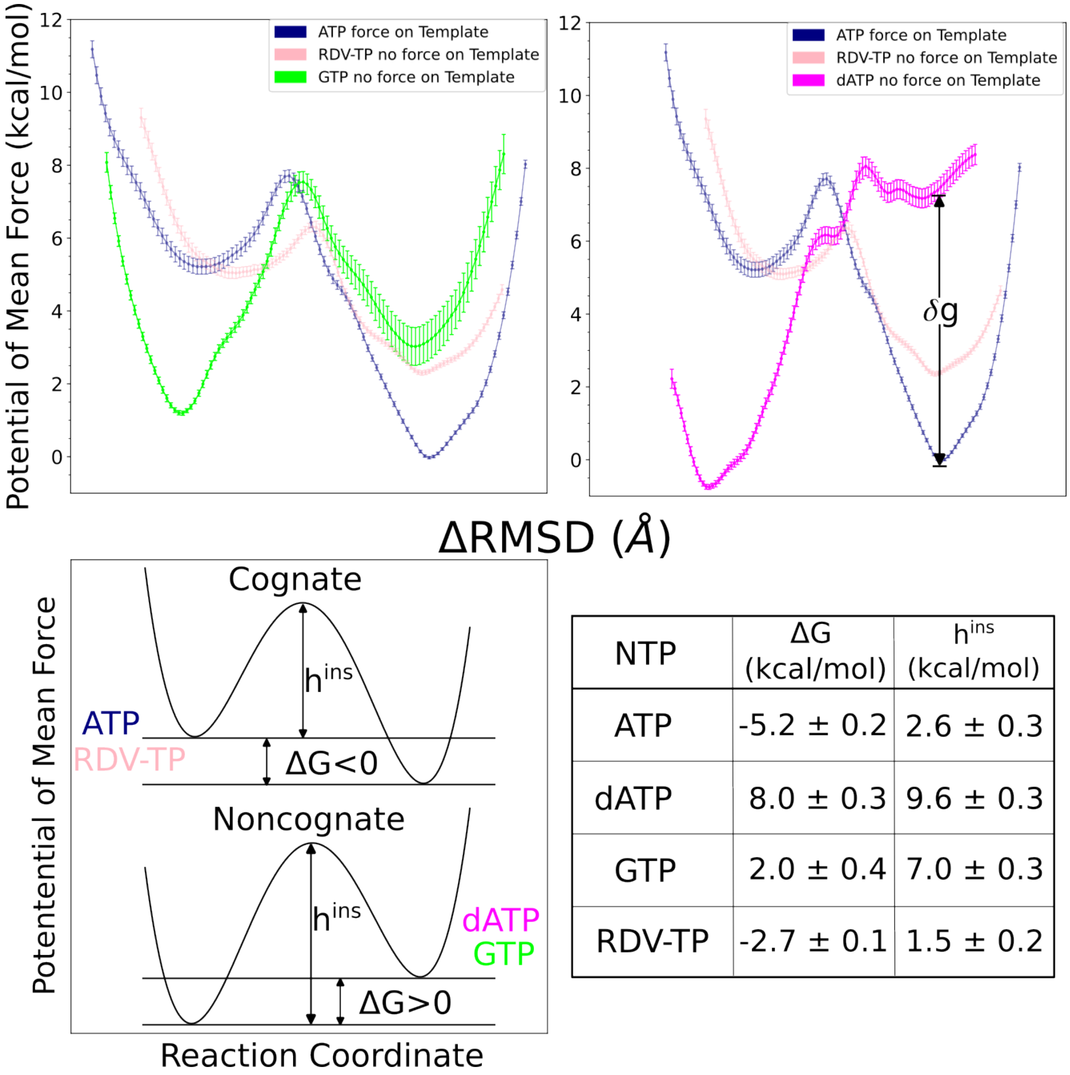
Figure 5.2: The potentials of mean force (PMFs) calculated for various NTP from initial binding (active site open) to the insertion (closed) state via umbrella sampling simulations. The difference of RMSDs with respect to open and closed reference structures $\delta RMSD = RMSD(X, X_{\text{Open ref}}) - RMSD(X, X_{\text{Closed ref}})$, was used as the reaction coordinate in the PMF construction. The upper left panel shows the PMFs for GTP, with (dark green) and without (light green) force on the template +1 nucleotide. In both cases, PMFs are shown in comparison with the PMFs obtained for cognate ATP (blue) and drug RDV-TP (pink). The upper right panel shows the PMFs for dATP, with (dark purple) and without (magenta) force on the template +1 nucleotide. The lower left and lower right panels display convergence plots of PMFs for GTP and dATP, respectively, in current umbrella sampling simulations, without force implemented to the template +1 nucleotide.

5.3.1 Constructing the PMFs for noncognate GTP/dATP from initial binding to insertion

Upon conformational samplings from the equilibrium ensemble simulations, we noticed essential variations of protein structural motifs along with diverse NTP dynamical responses. Accordingly, we included structural motifs and NTP configurations into constructing a collective reaction coordinate, based on RMSD changes with respective to the open and closed state structures (see 5.2).[50;96] We also selected appropriate reference states for each NTP initial binding/insertion system, and then proceeded to calculate the free energy profiles or potentials of mean force (PMFs) of NTP insertion and to quantify the processes with NTP selectivity from initial binding (active site open) to the insertion (closed).

From the constructed PMF, GTP upon initial binding displays notable stabilization in comparison with the insertion state, with $\Delta G = G_{\text{insertion}} - G_{\text{initial binding}}$ ~2 kcal/mol (> 0). This is in contrast with cognate ATP and RDV-TP analog insertion, which are more stabilized in the insertion state, demonstrating ΔG (< 0) values of -5.2 kcal/mol and -2.7 kcal/mol, respectively (Fig. 5.3).[50] Nevertheless, the initially bound GTP forms no WB base pairing with the template nt+1, whereas approximately 81% WB base pairing between GTP and the template is identified in the insertion state. Hence, there have to be other interactions to stabilize the initial binding GTP (to be addressed in next subsection). Additionally, the PMF calculations show that the non-cognate GTP is subject to an insertion barrier h^{ins} ~7 kcal/mol from the initial binding state, much larger than that of ATP and RDV-TP (h^{ins} ~2.6 kcal/mol and 1.5 kcal/mol, respectively; see Fig. 5.3).[50] Note that the convergence of the PMF for GTP was reached after 50 ns per window in the umbrella sampling simulation, following the protocol without (or with) enforcing on the Uracil template +1 nt (Fig. 5.2)

The placement of the GTP insertion PMF relative to that of ATP/RDV-TP was conducted according to alchemical simulations performed recently.[145] Given that the GTP-template WB base pairing geometries are stable in the insertion state (Fig. 5.3 *lower right*), we placed the GTP insertion state ~3 (2.95 ± 0.66) kcal/mol above that of the ATP in-



NTP	ΔG (kcal/mol)	h^{ins} (kcal/mol)
ATP	-5.2 ± 0.2	2.6 ± 0.3
dATP	8.0 ± 0.3	9.6 ± 0.3
GTP	2.0 ± 0.4	7.0 ± 0.3
RDV-TP	-2.7 ± 0.1	1.5 ± 0.2

Figure 5.3: The potentials of mean force (PMFs) calculated for NTP from initial binding (active site open) to the insertion (closed) state via umbrella sampling simulations. The difference of RMSDs with respect to open and closed reference structures $RMSD(X, X_{\text{Open Ref}}) - RMSD(X, X_{\text{Closed Ref}})$,[50] was used as the reaction coordinate in the PMF construction (see Methods). Top left shows the PMF for dATP (magenta) and top right for that of GTP (green), both in comparison with the PMFs obtained for cognate ATP (blue) and analog RDV-TP (pink).[50;145] Smoothing has been applied to each PMF for clarity (see original PMFs and converging tests in Fig. 5.2). Note that the placement of the PMF of GTP relative to that of ATP/RDV-TP is according to the alchemical calculation in the closed state[145] while the placement of the PMF of dATP relative to that of ATP/RDV-TP is still uncertain, while the relative binding free energy at the closed state, denoted as δg, is estimated between 2-7 kcal/mol (see text). Bottom left shows a summary of the different PMF profiles between cognate and noncognate insertions. Bottom right insertion free energy and barrier heights for each PMF shown on the top plots.

sertion state, as the alchemical calculations indicate that $\Delta\Delta G_{\text{binding}}$ ~3 kcal/mol for the relative binding free energy of GTP with respect to ATP in the insertion state.[145]

The dATP insertion PMF demonstrates even more stability in the initial binding state, with ΔG ~8.0 kcal/mol comparing with the insertion state. Meanwhile, the insertion state of dATP is only marginally stabilized. Besides, the non-cognate dATP is subject to an insertion barrier h^{ins} ~9.6 kcal/mol, the highest among all NTP insertion cases examined (Fig. 5.3). Similar to GTP, dATP does not form WC base pairing with the template nt+1 in the initial binding state but is capable of forming the WC pairing with the template, though intermittently (64%), in the insertion state. Therefore, there must also be some additional interactions stabilize or trapping dATP in the initial binding state, as reflected from the highly tilted PMF toward the initial binding state. Similar to GTP case, the constructed PMF of dATP reached convergence after 50 ns per window of the umbrella sampling simulation, following the protocol without (or with) force applied on the Uracil template (Fig. 5.2).

The placement of the dATP insertion PMF relative to that of ATP/RDV-TP is, however, less certain. Given a wide range of association configurations between dATP and template in the insertion state (Fig. 4.3 *upper right*), one cannot use the alchemical calculation results, which were calculated around local configurational space (for stabilized ATP and slightly destabilized dATP) with limited sampling.[145] An estimation is nevertheless made here, based on relative binding free energy calculated locally between dATP and cognate ATP (~2 kcal/mol) along with that from the mmPBSA calculation (~7 kcal/mol),[145] suggesting a range of free energetic values 2-7 kcal/mol in between the inserted dATP and ATP (shown as parameter δg in Fig. 5.3).

In order to further test the consistency of the PMF of insertion results for both GTP and dATP, we used SMD simulations to probe the stability of the insertion states for GTP (Fig. 5.4) and dATP (Fig. 5.5), respectively. To do that, the non-cognate GTP or dATP starting from the insertion state is pulled slightly away from the active site in the SMD simulations

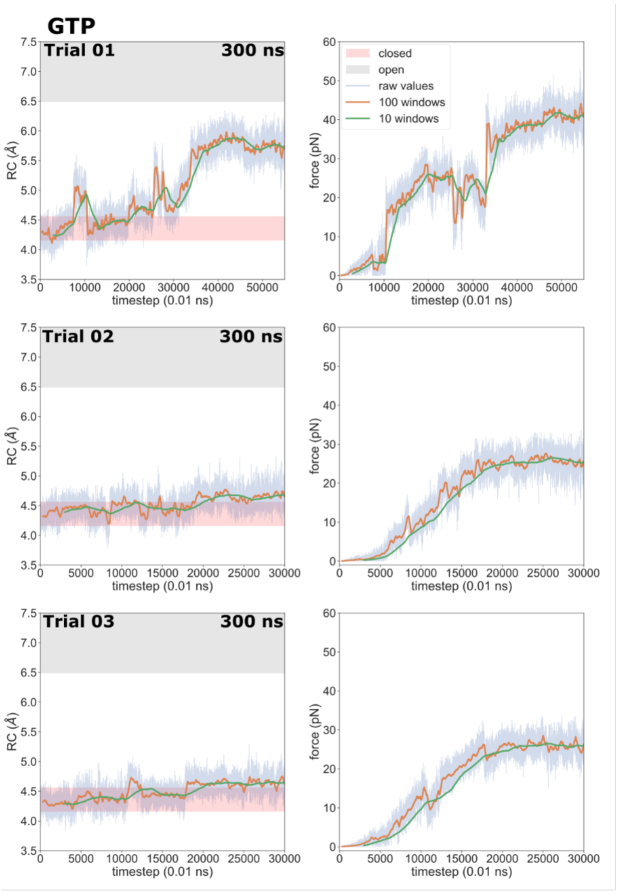


Figure 5.4: Results from steered MD (SMD) pulling GTP from the insertion (active-site closed) state well towards the initial binding (open) state at a rate of 1 $\mathring{A}$/ns (force constant 2.4 $\frac{kcal}{mol\mathring{A}^2}$). The reaction coordinate (RC) in pulling simulations, defined as the distance between the GTP and the active center (the center of mass of all C within 10 Å of the 3' RNA primer), is shown on the left panels. Instantaneous force applied in the SMD simulations is shown on the right panels. Raw data values, 100 window smoothed curves, and 10 window smoothed curves are drawn. ±1 standard deviation of the RCs in the open and closed wells from umbrella sampling are included (as the gray and pink bars on the left panels). Trial simulation 01 was run to a total of 550 ns. The other two trials 02 and 03 were run to a total of 300 ns.

(see Methods), i.e., to start from the insertion state (active site closed) to the initial binding state (open) under the SMD force. In the case of GTP, it was robustly maintained within the insertion state without being able to cross the barrier (from closed to open) from three trials of the SMD simulations (one 500 ns, two 300 ns). In contrast, dATP was able to be readily pulled from the insertion state toward the initial binding state (closed to open) under the

SMD force in all three simulations (300 ns each), demonstrating much lower stability or barrier from closed to open than that of GTP. These observations further support the instability of the dATP insertion state in comparison with the GTP insertion state, as being reflected from the PMFs constructed (Fig. 5.3). Next, we would proceed to examine additional interactions that stabilize or trap the non-cognate dATP or GTP in their initial binding state, i.e., as revealed from the PMFs constructed from the umbrella sampling simulation. How-

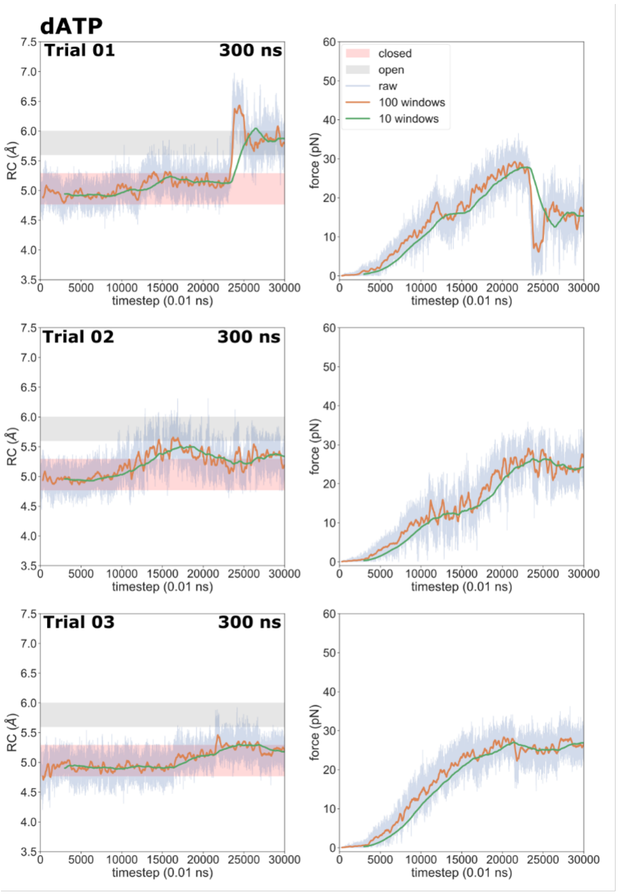


Figure 5.5: Results from steered MD (SMD) pulling dATP from the insertion (active-site closed) state well towards the initial binding (open) state at a rate of 1 $\mathring{A}$/ns (force constant 2.4 $\frac{kcal}{mol\mathring{A}^2}$). The reaction coordinate (RC) in pulling simulations, defined as the distance between the dATP and the active center (the center of mass of all C within 10 Å of the 3' RNA primer), is shown on the left panels. Instantaneous force applied in the SMD simulations is shown on the right panels. Raw data values, 100 window smoothed curves, and 10 window smoothed curves are drawn. ±1 standard deviation of the RCs in the open and closed wells from umbrella sampling are included (as the gray and pink bars on the left panels). All trial simulations were run to a total of 300 ns.

ever, before that, we want to examine whether the conformational space sampled between dATP/GTP and the template nt+1 near the initial binding (open) and insertion (closed) equilibrium in the umbrella sampling simulations overlap well with that from the equilibrium ensemble simulations. To do that, the NTP-template +1 nt base pairing geometries are compared between the equilibrium ensemble and the umbrella sampling simulations, from the latter three umbrella windows around the open/closed equilibrium were selected (40 ns RDV-TP/GTP/dATP and 160 ns ATP of simulation time per window). In Fig. 5.6, one can see that the sampled geometries from the umbrella samplings are comparatively restricted, especially in the NTP initial binding state, but overlap well with the dominant configurations sampled from the equilibrium ensemble simulations, in particular, in the insertion state. For ATP and GTP, the initial binding configurations from the umbrella samplings overlap well with some stabilized population from the equilibrium ensemble, though deviations from the equilibrium ensemble show in the umbrella sampling case, indicating potentially the forcing impacts from the umbrella sampling simulations. For RDV-TP initial binding, the stable configuration of the RDV-TP-template in base stacking is well sampled, as the umbrella sampling path was launched from the base stacking configurations.[50] For the dATP initial binding, the umbrella sampling simulation also well covers one stabilized population. Hence, the significantly stabilized state of GTP/dATP upon initial binding, detected from the umbrella sampling simulations, appears to be well located within a subspace in the equilibrium ensemble.

Since the initially bound non-cognate dATP or GTP could not be substantially stabilized by association with the template +1 nt (nor 3' end primer), it must be interactions from the RdRp protein along with the RNA scaffold around the active site that strongly hold the non-cognate dATP or GTP, which we examine and elaborate below.

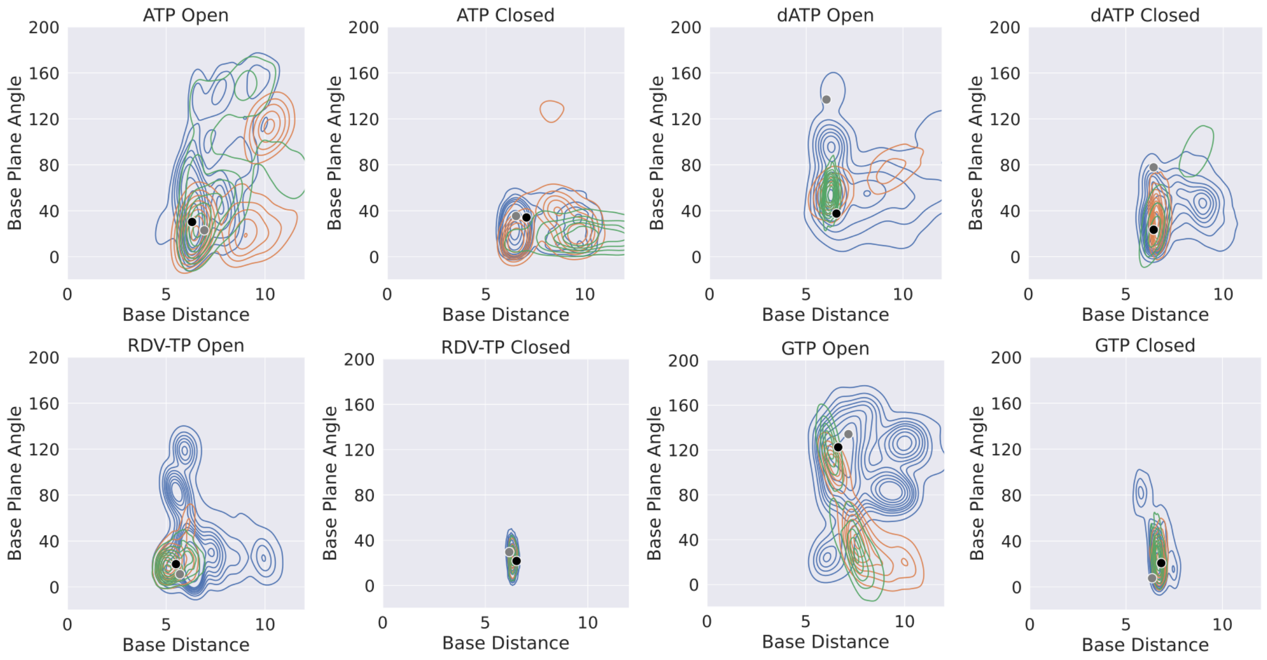


Figure 5.6: The NTP-template association geometry distributions obtained from the umbrella sampling simulations (for PMF calculations) in comparison with that from ensemble equilibrium simulations for various NTP species. A kernel density estimate has been used to visualize the data. Each simulation system open to closed, for ATP, RDV-TP, dATP and GTP. The equilibrium ensemble distribution is shown (blue) along with that obtained from the umbrella sampling (w/ force on template in orange; w/out force on template in green). The black dot indicates the reference state used to generate the initial paths for the umbrella sampling, and grey dot the reference state used in the alchemical calculations[145]

5.3.2 Trapping noncognate dATP/GTP upon initial binding by persistent HB interactions from motif A/F/G, to NTP, template nt +1, or 3' end primer.

Though there were no WC or WB base pairing interactions observed for non-cognate dATP/GTP upon initial binding toward the open active site, some populations of dATP/GTP are strongly stabilized upon the initial binding according to their insertion free energetics or PMFs (shown in Fig. 5.3). To gain understanding of this phenomena, we analyzed all HB interactions present among protein residues, NTP, and RNA strands (template and primer), around the active site. Given the variations amongst NTPs, we simplified the analyses by summing up overall HB populations exceeding 20% of occupancy (during the simulation 40ns/window) amongst the protein (residues within 10 Å of the active site center), template

+1 nt, 3' end of the primer, and the initially bound NTP; those HBs were further grouped according to interactions with the NTP on polyphosphate, sugar, or base (see Fig. 5.7). The cumulative HB measure is then normalized over that of the cognate ATP.

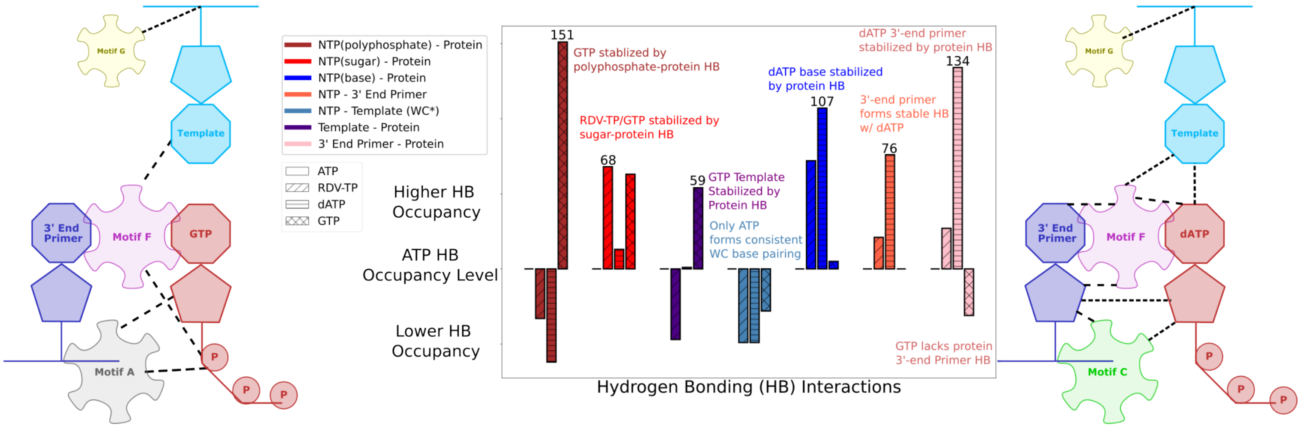


Figure 5.7: Summary of hydrogen bonding (HB) interactions that stabilize non-cognate GTP/dATP or surroundings upon initial binding from umbrella sampling simulations. Four interacting partners are considered: protein, incoming NTP (ATP, RDV-TP, dATP, and GTP), uracil template nucleotide +1, and the 3'-end primer. *Center:* The HBs with occupancy >20% in the simulations were identified among these four interaction partners. The relative HB occupancy levels are shown for the initial binding dATP and GTP (along with RDV-TP) with respect to that of cognate ATP as reference (with bar; see **SI Figure S10** for full HB statistics for all the simulation systems). Notable HB interactions that stabilize the non-cognate GTP and dATP initial binding systems are particularly denoted, respectively. Schematic and cartoon of key motifs stabilizing GTP (*left*) and dATP (*right*), along with template and 3'-end primer.

Notably, one finds that GTP initial binding is stabilized predominantly by HBs (and salt bridges in the case of positively charged residues LYS/ARG) formed between its polyphosphate group and the protein residues in motif's A and F. In addition, the protein residues (motif F and G) stabilize particularly the template nt +1, due to the absence of WC or WB base pairing between the initially bound GTP and the template nt. Meanwhile, there is lack of protein HB association with the 3' end primer around the initially bound GTP. However, this association is present around the initially bound ATP or RDV-TP, and the association appears even stronger around the non-cognate dATP initially bound.

In the case of dATP initial binding, though it does not form WC base pairing with the

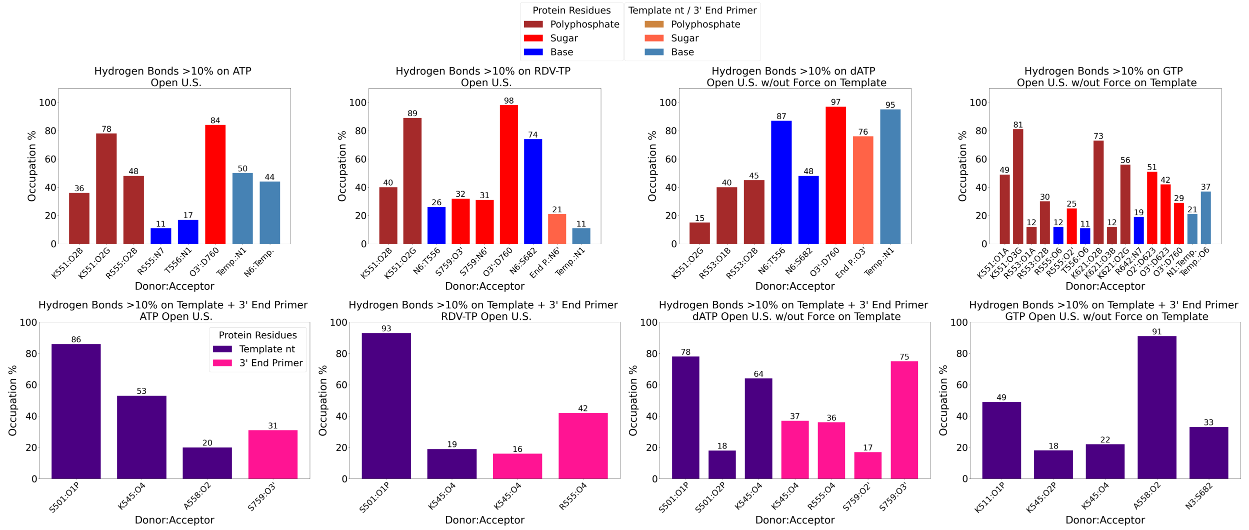


Figure 5.8: The hydrogen bond (HB) occupancy in umbrella sampling trajectories representing the initial binding (open) state minima for each PMF. The top row shows the HB for each NTP from protein/template +1 nt/3'-end primer. The bottom row shows the protein HB on the template +1 nt/3'-end primer. For each NTP (ATP, RDV-TP, dATP, and GTP), only unique interactions with a population greater than 10% are considered. The same color code used in Figure S2 and S3 is used to represent the different types of interactions.

template nt+1, the adenine base is nonetheless stabilized via protein HB (again from motif-F). Intriguingly, despite dATP's lack of a 2' OH group, it still maintains a strong HB via the sugar of the 3' end primer. As mentioned above, the 3' end primer around the initially bound dATP displays the strongest HB interactions among all the NTP's with the protein. Below, we show the individual HB interactions structurally and compare them with those in the case of cognate ATP or RDV-TP binding (Fig. 5.9 and Fig. 5.10). One can find statistics of HBs formed between NTP (polyphosphate, sugar, or base) and protein residues or template nt +1/3' end primer, and between protein residues and template nt+1 or the 3' end primer in Fig. 5.8.

Notably, we have found that upon initial binding, GTP exhibits very strong HB or salt bridge interactions between motif-F K551/R553 (together with motif-A K621) and the polyphosphate (see Fig. 5.9A). We attribute such interactions to hinder the insertion of GTP, or say, the phosphate-K551/R553 interactions contribute significantly to the barrier of GTP insertion. In the cognate initial binding, the ATP sugar forms a HB with motif-C D760. In contrast, GTP initial binding is mainly stabilized by HB between sugar and motif-A D623 (from umbrella sampling simulations). Instead of WB pairing with the non-cognate GTP upon initial binding, the template nt+1 base forms HBs with motif-F K545 and A558 (see Fig. 5.9B). Furthermore, the template nt +1 backbone forms HB with motif-G K511, as opposed to motif-G S501 seen in the other NTP binding cases. Overall, the protein motifs F and G seem to well stabilize the template nt +1 upon the base mismatched GTP binding.

In the case of RDV-TP initial binding via base stacking with the template nt +1 (in the absence of force on the template nt), however, only one HB is observed on the polyphosphate from motif-F K551 (see Fig. 5.9C), so that the RDV-TP won't be hindered by the phosphate interaction for its insertion.[50] In addition, the base stacking configuration of RDV-TP with the template allows for a unique HB to form between motif-B S682 and the RDV-TP base, while the template nt +1 has fewer HBs with motif-F than that upon GTP initial binding (Fig. 5.9D).

In comparison, as from current umbrella sampling simulations, dATP upon initial binding forms a single HB with template nt +1 at a very high occupancy of 95%. In addition, a HB is uniquely established between the dATP base and motif-F T556 at an occupancy of 87% (Fig. 5.10A). The dATP initial binding also forms two persistent HBs between the sugar and motif-C D760 and the 3' end primer, respectively. The template nt +1 is further stabilized by interactions with motif-F K545 and S501G, as seen with the cognate initial binding (Fig. 5.10B). Importantly, 3' end primer forms the most persistent HBs with motif-F K545/R555 and motif-C S759 in the case of dATP initial binding.

In the case of ATP initial binding, its base is stabilized by forming variable WC base pairing. Note that the initially bound ATP formed two HBs with template nt +1, with occupancies of 50% and 44%, respectively. The ATP sugar forms a consistent single HB with motif-C D760. Stable associations also form between motif-F K551/R553/R555 and the ATP polyphosphate (Fig. 5.10C). Such interactions were suggested to facilitate the cognate ATP insertion instead.[50] The template nt +1 forms stable HB with motif-F K545 and motif-G S501, similar to dATP initial binding. In addition, the 3' end primer forms only a single transient HB with motif-C S759 (Fig. 5.10D), weaker than that is present between the 3' end and motif-C/F upon the dATP initial binding (Fig. 5.10B).

5.4 Discussion

In current work, we have focused on computationally probing from initial binding to the insertion and selectivity mechanisms of noncognate natural nucleotides to SARS-CoV-2 RdRp. The insertion step process involves subtle but still substantial conformational change of the RdRp pol domain (Fig. 5.1), leading to essentially an active site open or nucleotide initial binding state to the active site closed or insertion state,[24] with coordination of seven highly conserved structural motifs. In all NTP incorporation systems simulated, the fingers subdomain displays similar conformational flexibility as the palm subdomain, which moves closer to the finger's subdomain in the insertion state than in the initial NTP binding state (Fig. 4.1). In the insertion state of cognate ATP or RDV-TP, motif-B and C have similar conformational flexibility (via RMSD) demonstrated in the equilibrium simulation, while this feature is absent in noncognate GTP and dATP. Overall, the motifs respond differently for each incoming NTP studied. The equilibrium ensemble simulations showed generically that the insertion state sampled a restricted subspace between the NTP and template nt +1 over that of the initial binding state, which indeed accommodates a wide range of configurations (Fig. 4.3). Additionally, RNA template/primer nucleotides or protein residues around the active site forms more HBs with the NTP in the insertion state than in the initial binding state (Fig. 4.5). Due to time scale limit of equilibrium sampling, we probed the NTP insertion dynamics and calculated the corresponding insertion energetics using the umbrella sampling methods. The energetic profile or insertion PMF was constructed along a collective RC according to a difference of RMSDs between the modeled intermediate structure of the RdRp and the open and closed reference states, respectively. The essential atomic coordinates included those of backbone atoms from seven highly conserved structural motifs (A to G), which are crucial for recruiting nucleotide substrates with selectivity and supporting catalysis,[17] and heavy atoms on incoming NTP along with (or without) the template nt +1. While such a choice on enforcing on the template nt or not played some significant role in the PMF construction of cognate ATP and analogue RDV-TP (see Ch. 3),[50] it made little difference in the non-cognate dATP or GTP results, e.g., as observed from the base pairing geometry measured between NTP and template nt +1 (Fig. 5.6Fig. 5.8). The insertion barriers were not affected by the above choice for dATP or GTP either Fig. 5.2. In contrast with the insertion PMFs of cognate ATP and RDV-TP analogue that bias toward a more stabilized insertion state, we have found intriguingly that the initial binding states for ncNTPs (dATP and GTP currently) can be much more stabilized than their insertion state. In addition, the insertion barrier of ncNTP becomes very high (up to 7-10 kcal/mol), also in contrast with the marginally low insertion barriers of cognate ATP and RDV-TP (∼2 kcal/mol) identified previously.[50] Such free energetic calculations and structural dynamics examinations reveal intrinsic nucleotide selectivity of SARS-CoV-2 RdRp, i.e., to inhibit the insertion of ncNTPs to the active site by trapping the ncNTPs *off-path* upon initial binding to the peripheral of the RdRp active site.

5.4.1 Free energetics favor insertion of cognate NTPs but disfavor insertion of non-cognate NTPs

In previous work,[50] we calculated the insertion PMFs for ATP and RDV-TP, respectively. While the RDV-TP initial binding could form WC base pairing with the template nt+1, a more stabilized conformation was found for RDV-TP in base stacking with the template nt +1. Besides, the RDV-TP insertion barrier would become high (h^{ins} ∼5 kcal/mol) when the enforcing in the umbrella sampling simulation included the template nt +1. The striking feature was due to the enhanced HB (or salt-bridge from positively charged LYS/ARG) interactions between the motif-F residues (K551/R553/R555) and the RDV-TP polyphosphate, upon the enforcing on the template nt. Removing forcing on the template nt+1, i.e., allowing sufficient thermal fluctuations on the template, however, the motif-F interactions with the polyphosphate reduce, and the insertion barrier lowers to a marginal value (h^{ins} ∼1.5 kcal/mol).

Upon the cognate ATP binding, including the template nt +1 for enforcing in the umbrella sampling simulations nevertheless supports an insertion barrier h^{ins} as low as ~2.6 kcal/mol. The well controlled template nt+1 facilitated WC base pairing and supported enhanced HB interactions between the motif-F K551/R555 and polyphosphate. Consequently, it was suggested that the motif-F interactions with the phosphate facilitate insertion of the cognate ATP, while such interactions appear to inhibit the ncNTP insertion.[146] Regardless of the exact protocol, the insertion state was always more stable than the initial binding state for the cognate ATP and analogue RDV-TP, both of which are actively biased or recruited into the closed active site for catalytic incorporation to the 3'-end of primer, experiencing marginally high barriers of insertion to the active site.[50]

In contrast, upon the initial binding of ncNTP (dATP or GTP), a large configuration space of the ncNTP with respect to the template+1 nt and 3'-end primer was identified, and the PMF always tilts toward the initial binding state, i.e., biased or energetically stabilized upon initial binding of certain ncNTP configurations (Fig. 5.3 & Fig. 5.2). Additionally, the insertion barriers insertion become very large for noncognate GTP and dATP (h^{ins} ~7.0 and 9.6 kcal/mol, respectively). The stability bias toward the initial binding state and tremendously large barrier of insertion seem to trap the ncNTP upon initial binding at certain configurations (or the *off-path*), which would likely lead to dissociation of the ncNTPs from the active site or from the RdRp in the end.

Our current discoveries on such stabilization of the non-cognate substrate upon initial binding may seem counterintuitive. Commonly, high binding affinity of a ligand substrate to the receptor protein indicates a preference of the receptor to the substrate.[147;148] Such an idea prevails in the drug design which aims at identifying high-affinity binders. In the current viral RdRp system, however, the NAC proceeds with two pre-chemical steps: substrate initial binding and insertion. As shown in current study, a high substrate binding affinity at the first step may also contribute to high insertion barrier for the second step, which slows down the NAC or an enzymatic cycle. Hence, the ncNTP stabilization or trapping upon initial binding becomes an intriguing but effective strategy to deter or inhibit the non-cognate substrate from further incorporation.

5.4.2 Key residues from conserved motifs detect and stabilize the ncNTP and its surroundings at the initial binding off-path state

Current free energy calculations reveal that the noncognate GTP/dATP is more stabilized upon initial binding to the RdRp active site than in the insertion state, in contrast with cognate ATP/RDV-TP to be more stabilized in the insertion state. To explain notable stabilized configurations sampled for the ncNTP initial binding state from the umbrella sampling simulations, we identified a variety of HB interactions around the active binding site formed among NTP (base, sugar and phosphate), RNA strands (template and primer), and the conserved protein motifs (Fig. 5.7). To well explain the trapping mechanism, we can also compare current system with a previously studied RdRp from Enterovirus or EV,[24] which is structurally similar to SARS CoV-2 RdRp. In EV RdRp, NTP insertion to the active site is suggested via several steps: first the base recognition, next the ribose sugar recognition, and then followed by the active site open to closed conformational transition, accompanied by the palm subdomain (motifs A,B,C,D,E) closing. Below we connect current observations of NTP initial binding to those suggested steps.

In the case of GTP initial binding, the template nt +1 (uracil) fails to interact with the mismatch GTP in the absence of WC or WB base pairing. The template nt +1 and GTP cannot be mutually stabilized, hence the protein motif-F residues (K545/A558) respond by stabilizing the template base. Additionally, motif-G K511 forms HB with the template RNA backbone, instead of S501 in the ATP/RDV-TP system. Given the non-stabilized GTP base, the sugar is next checked by the protein via motif-A D623, which is prominent upon GTP initial binding (in the umbrella sampling ensemble). The D623 interaction brings motif-A closer to GTP and allows a unique HB (or salt bridge) from K621 to the polyphosphate. Substantial HBs (or salt bridges) with the polyphosphate come additionally from motif-F K551/R553, similarly as seen in RDV-TP (with force on template nt +1). In both cases of initial binding (GTP and RDV-TP with forcing on the template nt+1), the insertion barriers are high. Such observations support the previously proposed mechanism that the Lys/Arg interactions with phosphates inhibit the ncNTP insertion but facilitate cognate NTP insertion,[50] or say, the protein-NTP phosphate interactions play a significant role in nucleotide selectivity or fidelity control. Recent NMR experiments on the structurally similar PV RdRp suggest that interactions from charged residues in motif-F are an important fidelity checkpoint, as they allow the triphosphate to rearrange prior to catalysis.[107]

In the case of dATP binding, since dATP has the same base as the cognate ATP, it is capable of forming stable WC base pairing with the template nt to pass the base recognition checkpoint. Nevertheless, dATP fails on the sugar recognition. The sugar is unable to anchor in the active site due to missing the 2' OH functional group. Instead, the 3' OH group forms HB with motif-C D760, similar to ATP and RDV-TP. Meanwhile, a unique HB forms between the dATP sugar and 3' end primer HB. As a result, dATP base adopts a tilted conformation, which supports only a single persistent HB w/ template nt +1 (Fig. 5.10A). The dATP base is then further stabilized by motif-F T556. The missing WC base pairing between dATP and the template nt +1 is supplemented by another HB formed between the dATP base and motif-F K545. Additional stabilization of dATP initial binding comes from the HB formed between the 3' end primer from motif-C S759, and the 3'-end primer further forms HB with dATP sugar. The dATP sugar further forms HB with D760 from motif C. Hence, it appears that the 3'-end primer plays an important role in stabilizing dATP upon initial binding *off-path*.

Although the non-cognate dATP initial binding stability appears perplexing, prior crystal structure studies on the structurally similar PV RdRp have used dCTP to stall and resolve the RdRp structure in the open state,[16] supporting a stable binding configuration of dNTP binding to the viral RdRp. In addition, experimental work on the nsp14 exonuclease enzyme of SARS-CoV-2 has shown that for excision of an incorrect nt, the nt needs to have a proper RNA sugar, 2' and 3' OH functional group,[149] indicating that the enzyme may require some nucleotide selectivity to prevent dNTP's from chemical incorporation. Furthermore, recent experimental studies have shown that an elongation complex soaked in solution with dNTPs had no catalytic activity.[150] Michaelis–Menten kinetics also showed the SARS-CoV-2 RdRp selectivity of ATP over dATP is ∼1000 ($\frac{V_{max}}{K_m}$ =0.02 in dATP and 23 in ATP).[58] Similar trends were observed in PV *in vitro* biochemical studies with ∼117 discrimination factor $(\frac{k_{cat}}{K_M})_{\mathrm{CTP}}/(\frac{k_{cat}}{K_M})_{\mathrm{dCTP}}$.[13] Other computational works also tested the design of inhibitors with ribose sugar modifications or removal of the OH function groups entirely.[131;132]

5.5 Conclusions

To conclude, we have employed all-atom MD simulations and demonstrated intrinsic or natural nucleotide selectivity of SARS-CoV-2 RdRp, in which the ncNTP is well stabilized or trapped upon initial binding to certain off-path configurations, as the highly conserved structural motifs F/G/A/C of the viral RdRp form HBs with ncNTP, RNA template nt, and/or 3'-end RNA primer. Intrinsically, it is not the polymerase enzyme that determines a right/cognate or wrong/non-cognate nucleotide substrate in the template-based polymerization. The cognate or non-cognate NTP species are determined by the RNA template nt primarily via the WC base pairing, and additionally by the 3'-end primer via base stacking etc. With incoming ncNTP that is incapable of stabilizing the template counterpart or the 3'-end primer, the RdRp structural motifs sense such instability, and then takes over to select against the ncNTP. Presumably, ncNTP can be rejected in case of low binding affinity, i.e., upon initial binding to RdRp. However, it appears that in SARS-CoV-2 RdRp, the ncNTP can be particularly stabilized, say *off-path*, upon initial binding to certain configurations, so that to be prevented or inhibited from further insertion to the active site. Such mechanism of nucleotide selection seems to be well supported by the two-step binding and insertion processes, *pre-chemically*, in the single-subunit viral polymerase enzymes. Partial off-path initial binding and inhibition for insertion of ncNTPs were also suggested computationally for single-subunit T7 RNAP.[62;96]

Technically, we have employed the umbrella sampling simulations to characterize the slow NTP insertion dynamics that is accompanied by the open to closed conformational changes around the RdRp active site. As the slow pre-chemical conformational transition likely takes place over milliseconds, it becomes indispensable to enhance computational sampling while limiting artifacts to be introduced. In the simulation, we well manipulated collective atomic coordinates from all structural motifs along with key players of NTP/template. Nevertheless, how to identify the most essential coordinates is a continuous challenging issue.[151;152] For multiple NTP species incorporation, enhanced computational sampling would become even more demanding, considering that multiple reaction paths exist.[153] Further exploration and validation of our current studies would require substantial experimental studies. Resolving high-resolution structures of the SARS-CoV-2 RdRp complexes with the stabilized *off-path* initial binding configurations of non-cognate dATP or GTP, as being proposed in current computational work, would be highly anticipated.

Chapter 6

Future Perspectives:Other Enhanced Sampling Methods and Improvement of the RC

In this dissertation, I have presented and discussed the construction of the initial binding and insertion states of SARS-CoV-2 RdRp. To probe this vital fidelity checkpoint, I modeled the cognate ATP, drug analogue RDV-TP, and noncognate dATP and GTP and run atomistic molecular dynamics simulations. Furthermore, nonequilibrium methods such as TMD and U.S. were used to compute the free energy of insertion (see 1.3). Although the U.S. simulations seem to span a significant portion of the base pair geometry subspace of the equilibrium ensemble, we noted unaccounted regions in the initial binding state (see Fig. 5.6). The initial binding state for RDV-TP yielded two stable configurations, which resulted in markedly different PMFs (see 3.3). Therefore, while our current findings[50;145] shed light on the fidelity of RdRp, a comprehensive understanding might necessitate exploring other methods for future RdRps or for a more thorough exploration of the CoV-2 Rdrp. In this final chapter, I will briefly and concisely describe some of the methods I explored but was unable to complete."

6.1 Improving Sampling

6.1.1 Hamiltonian Replica Exchange

A method briefly explored is Hamiltonian replica exchange[154] (HREX), currently implemented in Plumed for use with biomacromolecules. HREX operates similarly to parallel tempering, where the whole system is heated and exchanged, but it employs 'selective' heating given that energy is extensive. The mathematical formulation for this is as follows:

$$P(q;\lambda) \propto e^{-\frac{U(q)}{\lambda T}} = e^{-\frac{U(q)/\lambda}{T}} \tag{6.1}$$

$$U(q) = U_1(q) + U_2(q) \tag{6.2}$$

$$P(q;\lambda) \propto e^{-\frac{U_1(q)+U_2(q)/\lambda}{T}} \tag{6.3}$$

In these equations, P denotes the probability of a configuration q given a scale factor λ, U represents the total potential energy of the system, and T is the temperature. The terms U1 and U2 denote the contributions to the total potential energy from two different sources.

The significant distinction with HREX is that it selectively heats the solute using a factor denoted as λ. This factor, ranging from 0 (no bias) to 1, proportionally scales each replica based on a specified range, affecting only non-bonding and torsional parameters. If the area of interest within the system is known, this scaling can be applied specifically to that region. In the context of our research, the region of interest would be the NTP ± the template nucleotide +1. This approach results in a single trajectory or replica with improved sampling.

The implementation of HREX in Plumed suggests that this method can be further combined with other biasing techniques such as metadynamics and Umbrella Sampling (U.S.) to enhance sampling.[154] After a review of the relevant literature, it was noted that most applications of this method target intrinsically disordered protein simulations. However, only two instances were found where HREX was combined with U.S.

The first instance revolved around studying the curvature and buckling of peripheral proteins along a membrane.[155] One of their tested peptides showed multiple rotational orientations along their defined reaction coordinate for U.S. The researchers, therefore, applied HREX across a single window for 750 ns using 12 replicas. This approach allowed them to include both orientations in their U.S. reaction coordinate.

The second study sought to determine binding free energies in the SAMPL8 dataset binding to the CB8 receptor.[156] The researchers combined both HREX and U.S., using a reaction coordinate that represented the instantaneous separation between the center of mass of the host and ligand. The RC ranged between -2 and +2nm in steps of 0.1nm for a total of 40 windows. Each of these windows then also used HREX, so total simulations would be $40 \cdot$ N number of replicas used in HREX. Although feasible for their small system (CB8 is only 128 Amino Acids), for larger systems such as the CoV-2 RdRp (comprising 932 residues), this approach quickly becomes excessively costly.

When discussing the combination of U.S. with HREX, the developers stated that using HREX in each window represents the ideal way to combine the methods. However, they further emphasized that HREX should be considered as a last resort method, to be used only after testing at least three other biasing methods. They recommended metadynamics as the best starting point[157].[158]

6.1.2 Metadynamics

Metadynamics is indeed a popular and relatively simple technique that has been developed to bias molecular dynamics (MD) simulations for the purpose of calculating free energy. The initial formulation of Metadynamics works by depositing Gaussian functions of a constant height along a RC or CV.[159] However, the current most used version is well-tempered Metadynamics,[46] this approach adapts the method by gradually reducing the heights of the gaussians to account for oversampling (currently implemented in plumed).

One advantage of Metadynamics over techniques like U.S. is its simplicity of use. While U.S. requires a pre-generated path (as described in section 1.3.2), metadynamics only requires an equilibrated starting point. This makes it a more accessible and less demanding method for investigating the free energy landscapes of molecular systems. In my attempts to use well-tempered Metadynamics for RDV-TP, dATP, and GTP, the results were not as consistent as I hoped. For RDV-TP (from initial binding) and dATP (from insertion) – the low barrier systems – and GTP (from insertion) – the high barrier system, I found that the molecular dynamics simulations did not explore as much of the reaction coordinate space as intended. The dATP and GTP primarily stayed within their starting states, failing to transition between states as expected. For RDV-TP, although the simulation seemed to sample a similar value in both regions, the resultant potential of mean force (PMF) showed a single well, suggesting a failure in adequately sampling the entire free energy landscape. These findings underscore that a reaction coordinate that works well for one biasing method may not necessarily be effective for another.

6.2 Improving Reaction Coordinate

The selection of a suitable RC is indeed a fundamental and often challenging aspect of performing enhanced sampling simulations, whether it's Metadynamics, Umbrella Sampling, or another method. The RC needs to encapsulate the essential degrees of freedom of the system that describe the process of interest and capture the most important changes in the system's behavior.

In the case of modeling the SARS-CoV-2 RdRp, one could consider an RC that takes into account the key conformational changes that occur during the transition between the open and closed states. This could involve, for example, measuring the distance (as a vector) between key structural motifs A-G or a subset (A, D, F) in the RdRp. These motifs have been shown to play a significant role in the conformational changes of the RdRp and are therefore likely to be crucial for describing the transition between the open and closed states.[10;24]

Moreover, changes in the coordination of Mg ions have also been identified as a distinguishing feature of the open vs. closed states.[10;24] This provides another promising candidate for an RC.

The ability to define multiple dimensions for RC in biased simulations constitutes a significant advantage of this methodology. An emerging technique involves the application of machine learning methods to the reaction coordinate, with the aim of either reducing the relevant motions or using a high-dimensional RC.[151] An example of this implementation is MCOLVAR,[152] available in plumed. To effectively utilize this module, it is crucial to have defined metastable states for the system under investigation. In absence of these, the module's efficacy is largely dependent on a Principal Component Analysis (PCA). Presently, tests primarily focus on a single particle in 3D space, and evaluations on even the simplest biological molecule (such as an alanine dipeptide) have yet to be performed. Nonetheless, as these methods continue to be refined and developed, they are poised to play a crucial role in the future of free energy calculations.

In conclusion, while the task of defining suitable RCs is a nontrivial one, it is critical for the success of enhanced sampling simulations. It requires a good understanding of the system's behavior, informed by structural and functional knowledge, and can significantly influence the insights gained from the simulations.

Bibliography

[1] Gorbalenya, A. E. et al. *Nature Microbiology 2020 5:5* **2020**, *5*, 668–674.

[2] Gorbalenya, A. E.; Enjuanes, L.; Ziebuhr, J.; Snijder, E. J. *Virus Research* **2006**, *117*, 17–37.

[3] Finkel, Y. et al. *Nature 2020 589:7840* **2020**, *589*, 125–130.

[4] V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. *Nature Reviews Microbiology 2020 19:3* **2020**, *19*, 155–170.

[5] Subissi, L.; Posthuma, C. C.; Collet, A.; Zevenhoven-Dobbe, J. C.; Gorbalenya, A. E.; Decroly, E.; Snijder, E. J.; Canard, B.; Imbert, I. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, E3900–E3909.

[6] Snijder, E. J.; Decroly, E.; Ziebuhr, J. *Advances in Virus Research*; Academic Press, 2016; Vol. 96; pp 59–126.

[7] Gao, Y. et al. *Science (New York, N.Y.)* **2020**,

[8] Wang, Q. et al. *Cell* **2020**, *182*, 417–428.

[9] Kokic, G.; Hillen, H. S.; Tegunov, D.; Dienemann, C.; Seitz, F.; Schmitzova, J.; Farnung, L.; Siewert, A.; Höbartner, C.; Cramer, P. *Nature Communications 2021 12:1* **2021**, *12*, 1–7.

[10] Malone, B. F. et al. *Nature* **2023**, *614*, 781–787.

[11] Yin, W. et al. *Science* **2020**, *368*, 1499–1504.

[12] Lehmann, K. C.; Gulyaeva, A.; Zevenhoven-Dobbe, J. C.; Janssen, G. M.; Ruben, M.; Overkleeft, H. S.; Van Veelen, P. A.; Samborskiy, D. V.; Kravchenko, A. A.; Leontovich, A. M.; Sidorov, I. A.; Snijder, E. J.; Posthuma, C. C.; Gorbalenya, A. E. *Nucleic Acids Research* **2015**, *43*, 8416–8434.

[13] Campagnola, G.; McDonald, S.; Beaucourt, S.; Vignuzzi, M.; Peersen, O. B. *Journal of Virology* **2015**, *89*, 275–286.

[14] Temiakov, D.; Patlan, V.; Anikin, M.; McAllister, W. T.; Yokoyama, S.; Vassylyev, D. G. *Cell* **2004**, *116*, 381–391.

[15] Schwinghammer, K.; Cheung, A. C.; Morozov, Y. I.; Agaronyan, K.; Temiakov, D.; Cramer, P. *Nature Structural & Molecular Biology 2013 20:11* **2013**, *20*, 1298–1303.

[16] Gong, P.; Peersen, O. B. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, 22505–22510.

[17] Poch, O.; Sauvaget, I.; Delarue, M.; Tordo, N. *The EMBO Journal* **1989**, *8*, 3867–3874.

[18] Bruenn, J. A. *Nucleic Acids Research* **2003**, *31*, 1821–1829.

[19] Steitz, T. A.; Steitz, J. A. *Proceedings of the National Academy of Sciences of the United States of America* **1993**, *90*, 6498–6502.

[20] Ouirane, K. B.; Boulard, Y.; Bressanelli, S. *Journal of Biological Chemistry* **2019**, *294*, 7573–7587.

[21] Sydow, J. F.; Cramer, P. *Current Opinion in Structural Biology* **2009**, *19*, 732–739.

[22] Yu, J. *Computational and Mathematical Biophysics* **2014**, *2*, 141–160.

[23] Yuzenkova, Y.; Bochkareva, A.; Tadigotla, V. R.; Roghanian, M.; Zorov, S.; Severinov, K.; Zenkin, N. *BMC Biology* **2010**, *8*, 54.

[24] Shu, B.; Gong, P. *Proceedings of the National Academy of Sciences* **2016**, *113*, E4005–E4014.

[25] Cameron, C. E.; Moustafa, I. M.; Arnold, J. J. *Enzymes*; Academic Press, 2016; Vol. 39; pp 293–323.

[26] Oweida, T. J.; Kim, H. S.; Donald, J. M.; Singh, A.; Yingling, Y. G. *Journal of Chemical Theory and Computation* **2021**, *17*, 1208–1217.

[27] Galindo-Murillo, R.; Robertson, J. C.; Zgarbová, M.; Šponer, J.; Otyepka, M.; Jurečka, P.; Cheatham, T. E. *Journal of Chemical Theory and Computation* **2016**, *12*, 4114–4127.

[28] Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M. J.; Wang, J.; Duan, Y.; Luo, R. *Journal of Physical Chemistry B* **2012**, *116*, 7999–8008.

[29] Inakollu, V. S.; Geerke, D. P.; Rowley, C. N.; Yu, H. Polarisable force fields: what do they add in biomolecular simulations? 2020.

[30] Boothroyd, S. et al. *Journal of Chemical Theory and Computation* **2023**, *19*, 3251–3275.

[31] Brooks, B.; Karplus, M. *Proceedings of the National Academy of Sciences of the United States of America* **1983**, *80*, 6571–6575.

[32] Bayly, C. I.; Merz, K. M.; Ferguson, D. M.; Cornell, W. D.; Fox, T.; Caldwell, J. W.; Kollman, P. A.; Cieplak, P.; Gould, I. R.; Spellmeyer, D. C. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.

[33] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. 2006.

[34] Lindorff-LaImproved side-chain torsion potentials for the Amber ff99SB protein force fieldrsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Structure, Function and Bioinformatics* **2010**, *78*, 1950–1958.

[35] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *Journal of Chemical Theory and Computation* **2015**, *11*, 3696–3713.

[36] Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. *Journal of Chemical Theory and Computation* **2020**, *16*, 528–552.

[37] Xiong, Y.; Shabane, P. S.; Onufriev, A. V. *ACS Omega* **2020**, *5*, 25087–25094.

[38] Price, D. J.; Brooks, C. L. *Journal of Chemical Physics* **2004**, *121*, 10096–10103.

[39] Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophysical Journal* **2007**, *92*, 3817–3829.

[40] Ivani, I. et al. *Nature Methods* **2015**, *13*, 55–58.

[41] Zgarbová, M.; Šponer, J.; Otyepka, M.; Cheatham, T. E.; Galindo-Murillo, R.; Jurečka, P. *Journal of Chemical Theory and Computation* **2015**, *11*, 5723–5736.

[42] Wang, J.; Sattar, A. K.; Wang, C. C.; Karam, J. D.; Konigsberg, W. H.; Steitz, T. A. *Cell* **1997**, *89*, 1087–1099.

[43] Anderson, J. A.; Lorenz, C. D.; Travesset, A. *Journal of Computational Physics* **2008**, *227*, 5342–5359.

[44] Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. 2009.

[45] Kästner, J. Umbrella sampling. 2011.

[46] Barducci, A.; Bussi, G.; Parrinello, M. *Physical Review Letters* **2008**, *100*, 1–4.

[47] Miao, Y.; Feher, V. A.; Mccammon, J. A. **2015**,

[48] Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. *Journal of Chemical Physics* **2019**, *151*, 70902.

[49] Torrie, G. M.; Valleau, J. P. *Journal of Computational Physics* **1977**, *23*, 187–199.

[50] Romero, M. E.; Long, C.; La Rocco, D.; Keerthi, A. M.; Xu, D.; Yu, J. *Molecular Systems Design and Engineering* **2021**, *6*.

[51] Jordheim, L. P.; Durantel, D.; Zoulim, F.; Dumontet, C. *Nature Publishing Group* **2013**, *12*, 447.

[52] Beigel, J. H. et al. *New England Journal of Medicine* **2020**, *383*, 1813–1826.

[53] Agostini, M. L. et al. *mBio* **2018**, *9*, 1–15.

[54] Shannon, A.; Le, N. T. T.; Selisko, B.; Eydoux, C.; Alvarez, K.; Guillemot, J. C.; Decroly, E.; Peersen, O.; Ferron, F.; Canard, B. *Antiviral Research* **2020**, *178*, 104793–104808.

[55] Warren, T. K. et al. *Nature* **2016**, *531*, 381–385.

[56] Sheahan, T. P. et al. *Science Translational Medicine* **2017**, *9*, 3653.

[57] Choy, K. T.; Wong, A. Y. L.; Kaewpreedee, P.; Sia, S. F.; Chen, D.; Hui, K. P. Y.; Chu, D. K. W.; Chan, M. C. W.; Cheung, P. P. H.; Huang, X.; Peiris, M.; Yen, H. L. *Antiviral Research* **2020**, *178*, 104786.

[58] Gordon, C. J.; Tchesnokov, E. P.; Woolner, E.; Perry, J. K.; Feng, J. Y.; Porter, D. P.;

Götte, M. *Journal of Biological Chemistry* **2020**, *295*, 6785–6797.

[59] Dangerfield, T. L.; Huang, N. Z.; Johnson, K. A. *iScience* **2020**, *23*, 101849.

[60] Pruijssers, A. J. et al. *Cell Reports* **2020**, *32*, 107940.

[61] Robson, F.; Khan, K. S.; Le, T. K.; Paris, C.; Demirbag, S.; Barfuss, P.; Rocchi, P.; Ng, W. L. *Molecular Cell* **2020**, *79*, 710–727.

[62] Long, C.; Romero, M. E.; La Rocco, D.; Yu, J. *Computational and Structural Biotechnology Journal* **2021**, *19*, 3339–3348.

[63] Brueckner, F.; Ortiz, J.; Cramer, P. *Current Opinion in Structural Biology* **2009**, *19*, 294–299.

[64] Elfiky, A. A. *Life Sciences* **2020**, *253*, 117592.

[65] Zhang, L.; Zhou, R. *J. Phys. Chem* **2020**, *2020*, 6962.

[66] Zhang, L.; Zhang, D.; Wang, X.; Yuan, C.; Li, Y.; Jia, X.; Gao, X.; Yen, H. L.; Cheung, P. P. H.; Huang, X. *Physical Chemistry Chemical Physics* **2021**, *23*, 5852–5863.

[67] Kirchdoerfer, R. N.; Ward, A. B. *Nature Communications* **2019**, *10*, 1–9.

[68] Hillen, H. S.; Kokic, G.; Farnung, L.; Dienemann, C.; Tegunov, D.; Cramer, P. *Nature* **2020**, *584*, 154–156.

[69] Cermakian, N.; Ikeda, T. M.; Miramontes, P.; Lang, B. F.; Gray, M. W.; Cedergren, R. *Journal of Molecular Evolution* **1997**, *45*, 671–681.

[70] Filée, J.; Forterre, P.; Sen-Lin, T.; Laurent, J. *Journal of Molecular Evolution* **2002**, *54*, 763–773.

[71] de Farias, S. T.; dos Santos, A. P.; Rêgo, T. G.; José, M. V. *Frontiers in Genetics* **2017**, *8*, 1–7.

[72] Šali, A.; Blundell, T. L. *Journal of Molecular Biology* **1993**, *234*, 779–815.

[73] Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. *Nucleic Acids Research* **2004**, *32*, 665–666.

[74] Olsson, M. H.; SØndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *Journal of Chemical*

Theory and Computation **2011**, *7*, 525–537.

[75] Wang, J.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049–1074.

[76] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

[77] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *Journal of Molecular Graphics and Modelling* **2006**, *25*, 247–260.

[78] Morris, G. M.; Ruth, H.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. *Journal of Computational Chemistry* **2009**, *30*, 2785–2791.

[79] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. *SoftwareX* **2015**, *1-2*, 19–25.

[80] Meagher, K. L.; Redman, L. T.; Carlson, H. A. *Journal of Computational Chemistry* **2003**, *24*, 1016–1025.

[81] Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

[82] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

[83] Bussi, G.; Donadio, D.; Parrinello, M. *Journal of Chemical Physics* **2007**, *126*, 014101.

[84] Berendsen, H. J.; Postma, J. P.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

[85] V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. *Nature Reviews Microbiology* **2021**, *19*, 155–170.

[86] Romano, M.; Ruggiero, A.; Squeglia, F.; Maga, G.; Berisio, R. *Cells* **2020**, *9*, 1267.

[87] Chen, J.; Malone, B.; Llewellyn, E.; Grasso, M.; Shelton, P. M.; Olinares, P. D. B.; Maruthi, K.; Eng, E. T.; Vatandaslar, H.; Chait, B. T.; Kapoor, T. M.; Darst, S. A.; Campbell, E. A. *Cell* **2020**, *182*, 1560–1573.

[88] Malone, B.; Chen, J.; Wang, Q.; Llewellyn, E.; Choi, Y. J.; Olinares, P. D. B.; Cao, X.;

Hernandez, C.; Eng, E. T.; Chait, B. T.; Shaw, D. E.; Landick, R.; Darst, S. A.; Campbell, E. A. *Proceedings of the National Academy of Sciences of the United States of America* **2021**, *118*, 2102516118.

[89] Lin, S. et al. *Nucleic Acids Research* **2021**, *49*, 5382–5392.

[90] Boehr, D. D.; Arnold, J. J.; Moustafa, I. M.; Cameron, C. E. *Structure, Dynamics, and Fidelity of RNA-Dependent RNA Polymerases*; 2014; pp 309–333.

[91] Johnson, K. A. Conformational coupling in DNA polymerase fidelity. 1993.

[92] Schlick, T.; Arora, K.; Beard, W. A.; Wilson, S. H. *Theoretical Chemistry Accounts* **2012**, *131*, 1–8.

[93] Anand, V. S.; Patel, S. S. *Journal of Biological Chemistry* **2006**, *281*, 35677–35685.

[94] Long, C.; Yu, J. *Entropy* **2018**, *20*, 306.

[95] Long, C.; Chao, E.; Da, L. T.; Yu, J. *Computational and Structural Biotechnology Journal* **2019**, *17*, 638–644.

[96] Long, C.; Chao, E.; Da, L. T.; Yu, J. *Nucleic Acids Research* **2019**, *47*, 4721–4735.

[97] You, W.; Tang, Z.; Chang, C.-E. A. *J Chem Theory Comput* **2019**, *15*, 2433–2443.

[98] Schlitter, J.; Engels, M.; Krüger, P. *Journal of Molecular Graphics* **1994**, *12*, 84–89.

[99] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021.

[100] Liao, Q. *Progress in Molecular Biology and Translational Science*; Elsevier B.V., 2020; Vol. 170; pp 177–213.

[101] Grossfield, A. WHAM - Grossfield Lab. `http://membrane.urmc.rochester.edu/?page_id=126`.

[102] Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*; Chapman and Hall/CRC, 1994.

[103] Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *Journal of Computational Chemistry* **2011**, *32*, 2319–2327.

[104] And, A. S. T.; Elcock*, A. H. *Journal of the American Chemical Society* **2004**, *126*,

2208–2214.

[105] Johnson, K. A. *Biochimica et Biophysica Acta - Proteins and Proteomics* **2010**, *1804*, 1041–1048.

[106] Smidansky, E. D.; Arnold, J. J.; Reynolds, S. L.; Cameron, C. E. *Biochemistry* **2011**, *50*, 5016–5032.

[107] Yang, X.; Liu, X.; Musser, D. M.; Moustafa, I. M.; Arnold, J. J.; Cameron, C. E.; Boehr, D. D. *Journal of Biological Chemistry* **2017**, *292*, 3810–3826.

[108] Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; Robertson, D. L. *Nature Reviews Microbiology 2021 19:7* **2021**, *19*, 409–424.

[109] Araf, Y.; Akter, F.; Tang, Y. d.; Fatemi, R.; Parvez, M. S. A.; Zheng, C.; Hossain, M. G. *Journal of Medical Virology* **2022**, *94*, 1825–1832.

[110] Pandey, P.; Rane, J. S.; Chatterjee, A.; Kumar, A.; Khan, R.; Prakash, A.; Ray, S. *Journal of Biomolecular Structure and Dynamics* **2020**, 1–11.

[111] Papageorgiou, A. C.; Mohsin, I. *Cells 2020, Vol. 9, Page 2343* **2020**, *9*, 2343.

[112] Almehdi, A. M.; Khoder, G.; Alchakee, A. S.; Alsayyid, A. T.; Sarg, N. H.; Soliman, S. S. *Infection* **2021**, *49*, 855–876.

[113] Ullrich, S.; Nitsche, C. The SARS-CoV-2 main protease as drug target. 2020.

[114] Jin, Z.; Wang, H.; Duan, Y.; Yang, H. *Biochemical and Biophysical Research Communications* **2021**, *538*, 63–71.

[115] Tarannum, H.; Rashmi, K.; Nandi, S. *Current Drug Targets* **2021**, *23*, 802–817.

[116] Magazine, N.; Zhang, T.; Wu, Y.; McGee, M. C.; Veggiani, G.; Huang, W. *Viruses 2022, Vol. 14, Page 640* **2022**, *14*, 640.

[117] Papanikolaou, V.; Chrysovergis, A.; Ragos, V.; Tsiambas, E.; Katsinis, S.; Manoli, A.; Papouliakos, S.; Roukas, D.; Mastronikolis, S.; Peschos, D.; Batistatou, A.; Kyrodimos, E.; Mastronikolis, N. *Gene* **2022**, *814*, 146134.

[118] Mótyán, J. A.; Mahdi, M.; Hoffka, G.; Tőzsér, J. *International Journal of Molecular*

Sciences 2022, Vol. 23, Page 3507 **2022**, *23*, 3507.

[119] Iketani, S.; Mohri, H.; Culbertson, B.; Hong, S. J.; Duan, Y.; Luck, M. I.; Annavajhala, M. K.; Guo, Y.; Sheng, Z.; Uhlemann, A. C.; Goff, S. P.; Sabo, Y.; Yang, H.; Chavez, A.; Ho, D. D. *Nature 2022 613:7944* **2022**, *613*, 558–564.

[120] Hu, Y.; Lewandowski, E. M.; Tan, H.; Zhang, X.; Morgan, R. T.; Zhang, X.; Jacobs, L. M. C.; Butler, S. G.; Gongora, M. V.; Choy, J.; Deng, X.; Chen, Y.; Wang, J. *bioRxiv* **2022**, 2022.06.28.497978.

[121] Ahn, D. G.; Choi, J. K.; Taylor, D. R.; Oh, J. W. *Archives of Virology* **2012**, *157*, 2095–2104.

[122] Sevajol, M.; Subissi, L.; Decroly, E.; Canard, B.; Imbert, I. *Virus Research* **2014**, *194*, 90–99.

[123] Zhu, W.; Chen, C. Z.; Gorshkov, K.; Xu, M.; Lo, D. C.; Zheng, W. *SLAS Discovery* **2020**, *25*, 1141–1151.

[124] Grellet, E.; L'Hôte, I.; Goulet, A.; Imbert, I. Replication of the coronavirus genome: A paradox among positive-strand RNA viruses. 2022; `https://doi.org/10.1016/j.jbc.2022.101923`.

[125] Gong, P. *Frontiers in Molecular Biosciences* **2022**, *8*, 822218.

[126] Bravo, J. P.; Dangerfield, T. L.; Taylor, D. W.; Johnson, K. A. *Molecular Cell* **2021**, *81*, 1548–1552.

[127] Peng, Q.; Peng, R.; Yuan, B.; Wang, M.; Zhao, J.; Fu, L.; Qi, J.; Shi, Y. *Innovation* **2021**, *2*.

[128] Naydenova, K.; Muir, K. W.; Wu, L. F.; Zhang, Z.; Coscia, F.; Peet, M. J.; Castro-Hartmann, P.; Qian, P.; Sader, K.; Dent, K.; Kimanius, D.; Sutherland, J. D.; Löwe, J.; Barford, D.; Russo, C. J. *Proceedings of the National Academy of Sciences of the United States of America* **2021**, *118*, e2021946118.

[129] Ahmed, S.; Mahtarin, R.; Islam, S.; Das, A.; Al Mamun, S.; Samina, A.; Ali, A.; Das, S.; Al Mamun, A.; Ahmed, S. *Journal of Biomolecular Structure and Dynamics*

2022, *40*, 11111–11124.

[130] Koulgi, S.; Jani, V.; Mallikarjunachari Uppuladinne, V. N.; Sonavane, U.; Joshi, R. *Journal of Biomolecular Structure and Dynamics* **2022**, *40*, 7230–7244.

[131] Parise, A.; Ciardullo, G.; Prejanò, M.; De La Lande, A.; Marino, T. *Journal of Chemical Information and Modeling* **2022**, *62*, 4916–4927.

[132] Li, Y.; Zhang, D.; Gao, X.; Wang, X.; Zhang, L. *Journal of Physical Chemistry Letters* **2022**, *13*, 4111–4118.

[133] Giannetti, M.; Mazzuca, C.; Ripani, G.; Palleschi, A. *Molecules* **2023**, *28*, 191.

[134] Bera, S. C.; Seifert, M.; Kirchdoerfer, R. N.; van Nies, P.; Wubulikasimu, Y.; Quack, S.; Papini, F. S.; Arnold, J. J.; Canard, B.; Cameron, C. E.; Depken, M.; Dulin, D. *Cell Reports* **2021**, *36*.

[135] Luo, X.; Xu, T.; Gao, X.; Zhang, L. *Chinese Journal of Chemical Physics* **2022**, *35*, 407–412.

[136] Kierzek, R.; Burkard, M. E.; Turner, D. H. *Biochemistry* **1999**, *38*, 14214–14223.

[137] Bonomi, M. et al. *Nature Methods* **2019**, *16*, 670–673.

[138] Parrinello, M.; Rahman, A. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

[139] Waskom, M. L. *Journal of Open Source Software* **2021**, *6*, 3021.

[140] Hunter, J. D. *Computing in Science and Engineering* **2007**, *9*, 90–95.

[141] Huang, J.; Brieba, L. G.; Sousa, R. *Biochemistry* **2000**, *39*, 11571–11580.

[142] Arnold, J. J.; Smidansky, E. D.; Moustafa, I. M.; Cameron, C. E. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **2012**, *1819*, 948–960.

[143] Sultana, S.; Solotchi, M.; Ramachandran, A.; Patel, S. S.; Wek, R. C. *Journal of Biological Chemistry* **2017**, *292*, 18145–18160.

[144] Lazim, R.; Suh, D.; Choi, S. Advances in molecular dynamics simulations and enhanced sampling methods for the study of protein systems. 2020; `https://www.mdpi.com/1422-0067/21/17/6339/htmhttps://www.mdpi.com/1422-0067/21/17/6339`.

[145] Chunhong Long,; Ernesto Romero, M.; Liqiang Dai,; Jin Yu, *Physical Chemistry*

Chemical Physics **2023**, *25*, 13508–13520.

[146] Shi, J.; Perryman, J. M.; Yang, X.; Liu, X.; Musser, D. M.; Boehr, A. K.; Moustafa, I. M.; Arnold, J. J.; Cameron, C. E.; Boehr, D. D. *Biochemistry* **2019**, *58*, 3735–3743.

[147] Alonso, H.; Bliznyuk, A. A.; Gready, J. E. *Medicinal Research Reviews* **2006**, *26*, 531–568.

[148] Schimunek, J. et al. *bioRxiv* **2023**,

[149] Jones, A. N.; Mourão, A.; Czarna, A.; Matsuda, A.; Fino, R.; Pyrc, K.; Sattler, M.; Popowicz, G. M. *Scientific Reports* **2022**, *12*.

[150] Petushkov, I.; Esyunina, D.; Kulbachinskiy, A. *FEBS Journal* **2023**, *290*, 80–92.

[151] Bhakat, S. *RSC Advances* **2022**, *12*, 25010–25024.

[152] Bonati, L.; Trizio, E.; Rizzi, A.; Parrinello, M. *bioRxiv* **2023**,

[153] Chong, L. T.; Saglam, A. S.; Zuckerman, D. M. Path-sampling strategies for simulating rare events in biomolecular systems. 2017.

[154] Bussi, G. *https://doi.org/10.1080/00268976.2013.824126* **2013**, *112*, 379–384.

[155] Stroh, K. S.; Risselada, H. J. *Journal of Chemical Theory and Computation* **2021**, *17*, 5276–5286.

[156] Markthaler, D.; Kraus, H.; Hansen, N. *Journal of Computer-Aided Molecular Design* **2022**, *36*, 1–9.

[157] PLUMED Masterclass 21-4.1 - YouTube. `https://www.youtube.com/watch?v=xAWhtFk5cMMhttps://www.youtube.com/watch?v=q0RHlFAk544`.

[158] PLUMED Masterclass 21-4.1 - YouTube. `https://www.youtube.com/watch?v=LexZoELjR5chttps://www.youtube.com/watch?v=q0RHlFAk544`.

[159] Laio, A.; Parrinello, M. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 12562–12566.

Appendix

Generating a Nucleotide Analogue Force Field

Below are the stepwise instructions for generating a force field for nucleotide analogue. In general it can be considered a specific case of a generating a force field for a small molecule or drug. Since it is a nucleotide analogue the tricky part comes in when selecting atom types. The first step is simply calculating the partial charges. **Note:** This is done on a truncated (at C5') nucleotide as triphosphate parameters are taken from already calculated values for AMBER.

Step 1: Generating Partial Charges + RESP Fitting

To generate partial charges we will use antechamber and Gaussian. After creating your structure with hydrogens run the following script:

```
db=RDV_w_H.pdb
antechamber -i $pdb -fi pdb -o $pdb".gin" -fo gcrt
-nc 0 -gk "# HF/6-31G* Opt SCF=Tight Pop=MK IOp(6/33=2, 6/41=10, 6/42=17)"
-gm "%mem=700MB" -gn "nproc=8"
```

The -i, -fi, -o, -fo specify the input and out file names + types. The rest are guassian arguments with -nc being charge, -gk gassuian input, and gm/gn resources requested. This will produce a file with .gin extension which can be run via Gaussian. After this completes you should have a log file which we will use for RESP fitting and generating our charges. Now we will perform resp fitting with a charge restraints on the O5'/O3' oxygens and their hydrogens: O5'(-.6223e),H5T (.4295e), O3'(-.6541e) and H3'(.4376e) Create a file titled restraints with information below:

```
CHARGE  0.4295 19 H5*
CHARGE -0.6223  1 O5*
CHARGE  0.4376 24 H3*

CHARGE -0.6541 6  O3*
```

The third column here is the index of the atoms listed above, the * is the same prime. ie O5*=O5'. **Critical: It is important that you select the O5'/O3' as they are indexed + named as in your structure.** After creating this file you run the following set of commands to run the RESP2 method.

```
log=../RDV_w_H
#Espgen generates contains ESP fitting centers created by gaussian
espgen -i $log".log" -o "RDV_w_H.esp"

# Create input structure file for RESP fitting
antechamber -i $log".log" -fi gout -o "RDV_w_H.ac" -fo ac

# Create RESP input files using respgen with restraints
respgen -i RDV_w_H.ac -o RDV_w_H.respin1 -f resp1 -a restraints.txt

# Create second RESP input
respgen -i RDV_w_H.ac -o RDV_w_H.respin2 -f resp2

# First Stage of RESP fitting
resp -O -i RDV_w_H.respin1 -o RDV_w_H.respout1 -e RDV_w_H.esp -t qout_stage1
 ↪  -q QIN

# Second Stage RESP Fitting, charges are printed to qout_stage1 file
resp -O -i RDV_w_H.respin2 -o RDV_w_H.respout2 -e RDV_w_H.esp -t qout_stage2
 ↪  -q qout_stage1
```

```
# Attach charges to a mol2 file using antechamber
antechamber -i RDV_w_H.ac -fi ac -o RDV_w_H.mol2 -fo mol2 -c rc -cf
 ↪  qout_stage2
# Attach charges to a mol2 file using PDB w/ Hydrogens as reference
antechamber -i RDV_w_H.pdb -fi pdb -o RDV_w_H_pdb.mol2 -fo mol2 -c rc -cf
 ↪  qout_stage2
```

At the end of these commands you will have a mol2 file with partial charges attached.

Step 2: Torsional Parameters

After selectin proper atom types comparing to the cognate or standard nucleoside, a force field file can be generated:

```
    mol2=RDV_w_H_pdb_addangles.mol2
    # Read in mol2 and convert to prepi format (used to make ff)
    # -j tells antechamber how to get bond info, -pf remove intermediate
     ↪  files
    # antechamber -i £mol2 -fi mol2 -o MPV_A.prepi -fo prepi -j 0 -pf n
    parmchk2 -f prepi -i RDV.prepi -o RDV.frcmod
```

This can be converted into any format for other packages.

Plumed Input Files

Target MD

Below are sample input files for TMD in the ATP case for both initial and forward paths.

Forward Path

```
 # Select refrence RMSD to match
rmsd: RMSD REFERENCE=ATP_closedref.pdb  TYPE=OPTIMAL

# Moving Restraint for target
MOVINGRESTRAINT ...
        ARG=rmsd
        AT0=0.0 STEP0=0      KAPPA0=0
        AT1=0.0    STEP1=50000000   KAPPA1=210000.0
... MOVINGRESTRAINT

PRINT ...
FILE=COLVAR ARG=rmsd
STRIDE=5000
... PRINT
ENDPLUMED
```

Backward Path

```
 # Select refrence RMSD to match
rmsd: RMSD REFERENCE=ATP_closedref.pdb  TYPE=OPTIMAL

# Moving Restraint for target
MOVINGRESTRAINT ...
        ARG=rmsd
        AT0=0.0 STEP0=0      KAPPA0=0

        AT1=0.0    STEP1=50000000   KAPPA1=210000.0

... MOVINGRESTRAINT


PRINT ...

FILE=COLVAR ARG=rmsd

STRIDE=5000

... PRINT

ENDPLUMED
```

Umbrella Sampling Example of an input for a singular window at RC=-0.09

```
RMSD_XA: RMSD REFERENCE=ATP_openref_TTP.pdb TYPE=OPTIMAL

RMSD_XB: RMSD REFERENCE=../inputs/ATP_closedref_TTP.pdb  TYPE=OPTIMAL


COMBINE  LABEL=deltaDrmsd ARG=RMSD_XA,RMSD_XB COEFFICIENTS=1,-1 POWERS=1,1 PERIODIC=NO


RESTRAINT ARG=deltaDrmsd AT=-0.09 KAPPA=210000.0 LABEL=restraint


PRINT ...

FILE=COLVAR ARG=RMSD_XA,RMSD_XB,deltaDrmsd

STRIDE=5000

... PRINT


ENDPLUMED
```

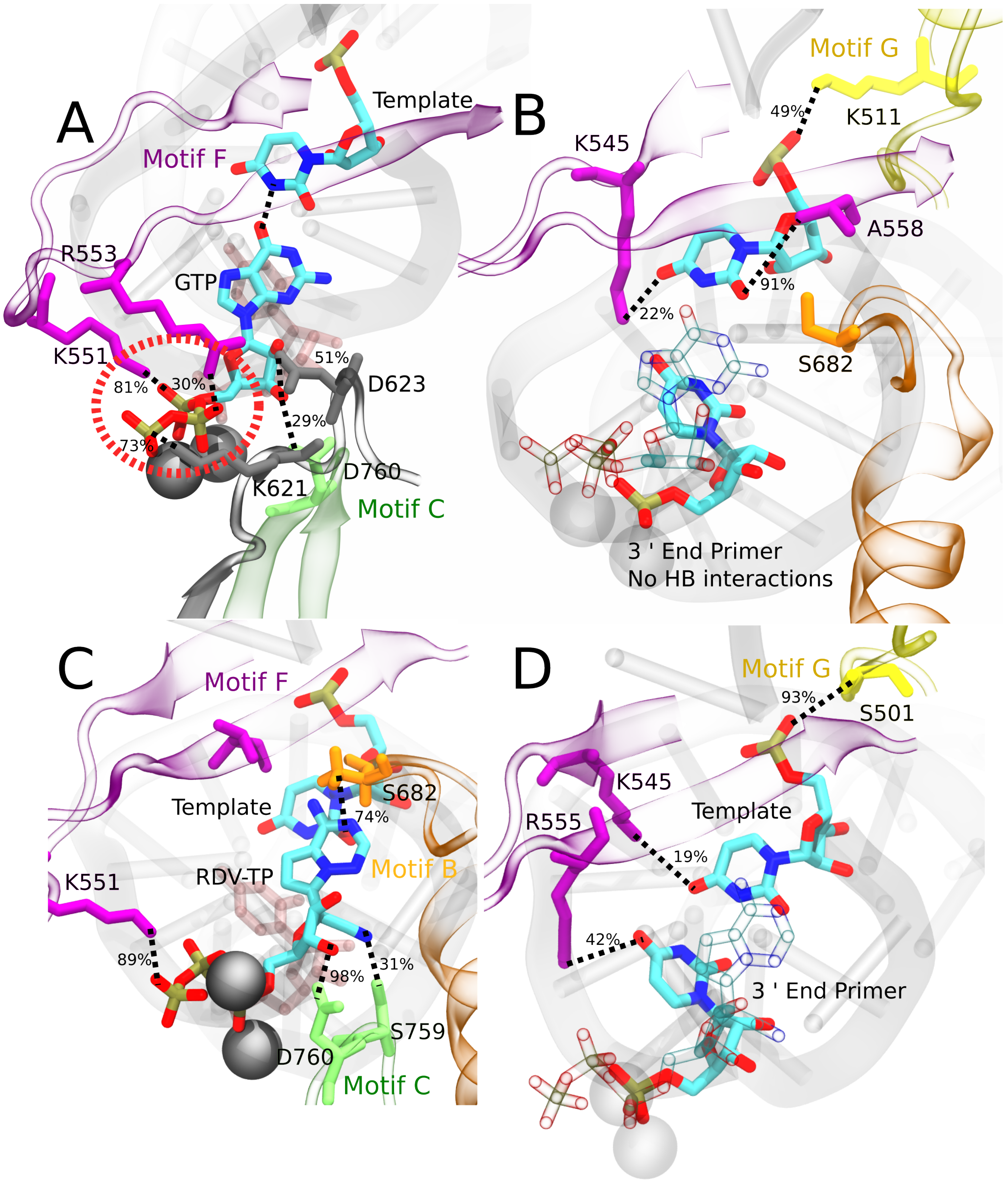


Figure 5.9: Comparing key interactions that stabilize the non-cognate GTP (and surroundings) and drug analogue RDV-TP in the initial binding system. The conserved protein motifs are shown in cartoon representation while interacting residues from these motifs are shown in licorice using the same color. A/C: Incoming GTP/RDV-TP and template nucleotide are shown in licorice colored by atom name. Hydrogen bonds (HBs) formed between GTP (or RDV-TP) and template nucleotide uracil/protein/3' end RNA primer. Orange circle highlights the essential interactions involved in 'trapping' non-cognate GTP phosphate in the initially bound state (A), which are absent for RDV-TP phosphate (C). B/D: Incoming GTP/RDV-TP is shown in transparent representation for clarity, 3'-end primer and template nucleotide are shown in licorice colored by atom name. HBs formed between the protein and template for stabilization, without proper WB base pairing for GTP-template (B) and with template base stacking in RDV-TP (D). For detailed HB occupancy plots see Fig. 5.8

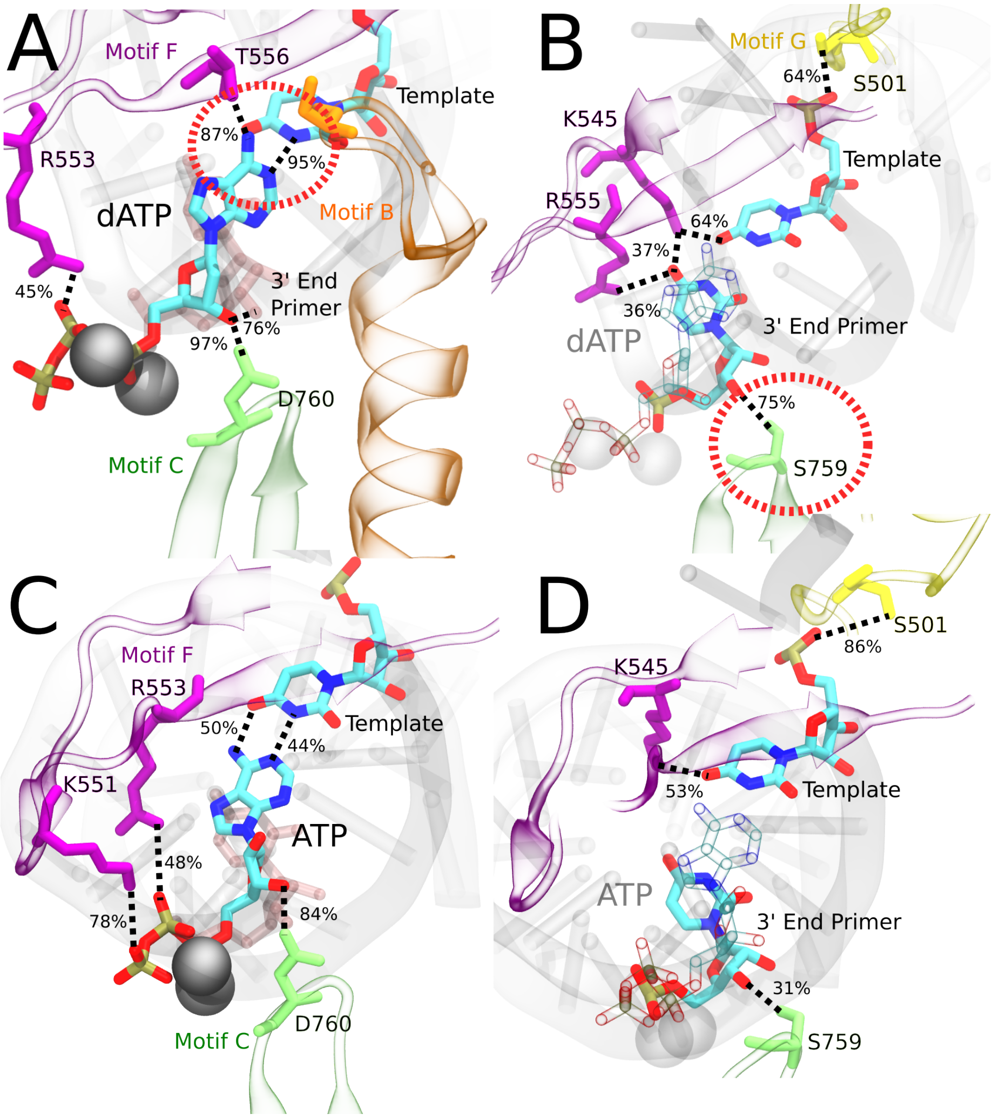


Figure 5.10: Comparing the key interactions that stabilize the non-cognate dATP (and surroundings) and cognate ATP in the initial binding system. The conserved protein motifs are shown in cartoon representation while interacting residues from these motifs are shown in licorice using the same color. A/C: Incoming dATP/ATP and template nucleotide are shown in licorice colored by atom name. Hydrogen bonds (HBs) formed between dATP (or ATP) and template nucleotide uracil/protein/3'-end RNA primer. Orange circle highlights the strongest interactions involved in 'trapping' non-cognate dATP in the initially bound state. B/D: Incoming dATP/ATP is shown in transparent representation for clarity, 3'-end primer and template nucleotide are shown in licorice colored by atom name. HBs formed between the protein and template nucleotide uracil / 3'-end RNA primer for stabilization, in the absence of proper WC base pairing in dATP (B). HBs formed between the protein and template nucleotide uracil / 3'-end RNA primer for stabilization, with proper WC base pairing in ATP (D). For detailed HB occupancy plots see Fig. 5.8.