Fall 12-16-2020

Computational Investigations of Biomolecular Mechanisms in Genomic Replication, Repair and Transcription

Thomas Dodd

Recommended Citation

Dodd, Thomas, "Computational Investigations of Biomolecular Mechanisms in Genomic Replication, Repair and Transcription." Dissertation, Georgia State University, 2020.
doi: https://doi.org/10.57709/20345880

COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MECHANISMS IN GENOMIC REPLICATION, REPAIR AND TRANSCRIPTION

by

THOMAS DODD

Under the Direction of Ivaylo Ivanov, Ph.D

ABSTRACT

High fidelity maintenance of the genome is imperative to ensuring stability and proliferation of cells. The genetic material (DNA) of a cell faces a constant barrage of metabolic and environmental assaults throughout the its lifetime, ultimately leading to DNA damage. Left unchecked, DNA damage can result in genomic instability, inviting a cascade of mutations that initiate cancer and other aging disorders. Thus, a large area of focus has been dedicated to understanding how DNA is damaged, repaired, expressed and replicated. At the heart of these processes lie complex macromolecular dynamics coupled with intricate protein-DNA interactions. Through advanced computational techniques it has become possible to probe these mechanisms at the atomic level, providing a physical basis to describe biomolecular phenomena. To this end, we have performed studies aimed at elucidating the dynamics and interactions intrinsic to the functionality of biomolecules critical to maintaining genomic integrity: modeling the DNA editing mechanism of DNA polymerase III, uncovering the DNA damage recognition/repair mechanism of thymine DNA glycosylase and linking genetic disease to the functional dynamics of the pre-initiation complex transcription machinery. Collectively, our results elucidate the dynamic interplay between proteins and DNA, further broadening our understanding of these complex processes involved with genomic maintenance.

INDEX WORDS: Computational biophysics, Computational chemistry, Molecular dynamics, DNA replication, DNA repair, Transcription

COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MECHANISMS IN GENOMIC REPLICATION, REPAIR AND TRANSCRIPTION

by

THOMAS DODD

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2020

COMPUTATIONAL INVESTIGATIONS OF BIOMOLECULAR MECHANISMS IN GENOMIC REPLICATION, REPAIR AND TRANSCRIPTION

by

THOMAS DODD

Committee Chair: Ivaylo Ivanov

Committee: Donald Hamelberg
Kathryn Grant

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
December 2020

DEDICATION

I would like to dedicate this work to my family and friends. Specifically, my loving fiancée Haley Jackson and my parents Terry and Tom Dodd. Without your love and support, I would have never made it this far. Thank you from the bottom of my heart.

ACKNOWLEDGEMENTS

First, I would like to acknowledge my advisor, Prof. Ivaylo Ivanov who has served as wonderful mentor. During my tenure he has never ceased to inspire me, as well as, drive my creative thinking. More importantly, he has spent countless hours teaching and molding me into an exceptional scientist. For that, I am forever grateful.

I have also been fortunate to have had some fantastic collaborations. Specifically, I would like to acknowledge Meindert Lamers, Margherita Botto and Rafael Fernandez-Leiro whose work is prominently featured in this dissertation. Your patience and guidance have been truly inspirational.

The Ivanov Group members (past and present) who I've had the pleasure of working with: Chunli Yan, Brad Kossmann, Kathleen Carter, Kurt Martin, Ashutosh Shandilya, Jina Yu and Bernard Scott. Thank you for your support and camaraderie.

I would also like to thank the members of my committee: Donald Hamelberg and Kathryn Grant. Thank you for your guidance and support during my graduate career and my dissertation preparation and defense.

Additionally, I would also like to thank the MBD for providing my funding and ample research opportunities.

Finally, I owe the most appreciation to everyone in my family: Haley Jackson, Terry, Tom, Katie, Bobby and Jenny Dodd, Catherine Callahan and Laura and Joey Caraway. Your love and support have been, and will continue to be, the driving force behind all of my endeavors. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. GENOMIC REPLICATION, REPAIR, TRANSCRIPTION AND KEY OUTCOMES

1.1 DNA Replication and Editing

Cellular chromosomal replicases are responsible for accurate and faithful DNA replication, which is essential for genomic stability. Replicases in all living organisms are tripartite and contain a DNA polymerase (Pol III in bacteria; Pol δ and ε in eukaryotes), a sliding clamp for processivity (β in bacteria, proliferating cell nuclear antigen in eukaryotes) and a clamp loader (DnaX in bacteria; replication factor C in eukaryotes)[1]. These proteins assemble at the origin of replication in the genome, along with a helicase (DnaB in bacteria; mini-chromosome maintenance complex in eukaryotes), a primase and single-stranded binding proteins to initiate the replication fork (Figure 1.1.1). At the replication fork, helicases unwind double-stranded DNA (dsDNA) into two complementary single strands of DNA (template strands). With the help of DNA polymerase and its processivity factor, nucleotides are added according to Watson-Crick pairing to the ssDNA to form two identical copies of the original dsDNA molecule. Due to the antiparallel nature of dsDNA, polymerases can only extend pre-existing primers from their 3’-OH end. Thus, one strand is synthesized continuously (leading strand) and the other in short ~1000 nucleotide sections called Okazaki fragments (lagging strand).

In recent years considerable progress has been made in the characterization of the chemical and structural description of DNA polymerases. Across all forms of life the structure of polymerases share a similar organization, and the nucleotidyl transferase reaction of adding new nucleotides appears conserved[2]. Structurally, all known DNA polymerases appear to resemble a right hand, where the functional domains are classified as fingers, palm and thumb domains. The catalytic reaction of adding new nucleotides occurs in the palm domain which is the most highly conserved subdomain. Here, the active site contains two invariable aspartic amino acids that coordinate two magnesium ions and form the basis of the two-metal ion catalytic mechanism, conserved in all polymerases studied to date[3].

Polymerases are highly accurate during the synthesis of DNA. In the rare instance that a mistake is made during replication, replicative polymerases have a 3’ - 5’ exonuclease to remove the incorrect nucleotide (termed editing). A functionally competent exonuclease is imperative to genomic stability as errors generated during synthesis, if left unchecked, can result in a cascade of mutations, some of which have been linked to cancer and premature aging[4]. The exonuclease domain is either built-in to the polymerase or located on a separate enzyme that associates with the polymerase. In either case, the active site of the exonuclease is distal to polymerase active site imposing strict structural and spatial requirements on the ssDNA transfer between sites[1, 6]. Despite decades of experimental work, the molecular details of this transfer mechanism had remained unknown.

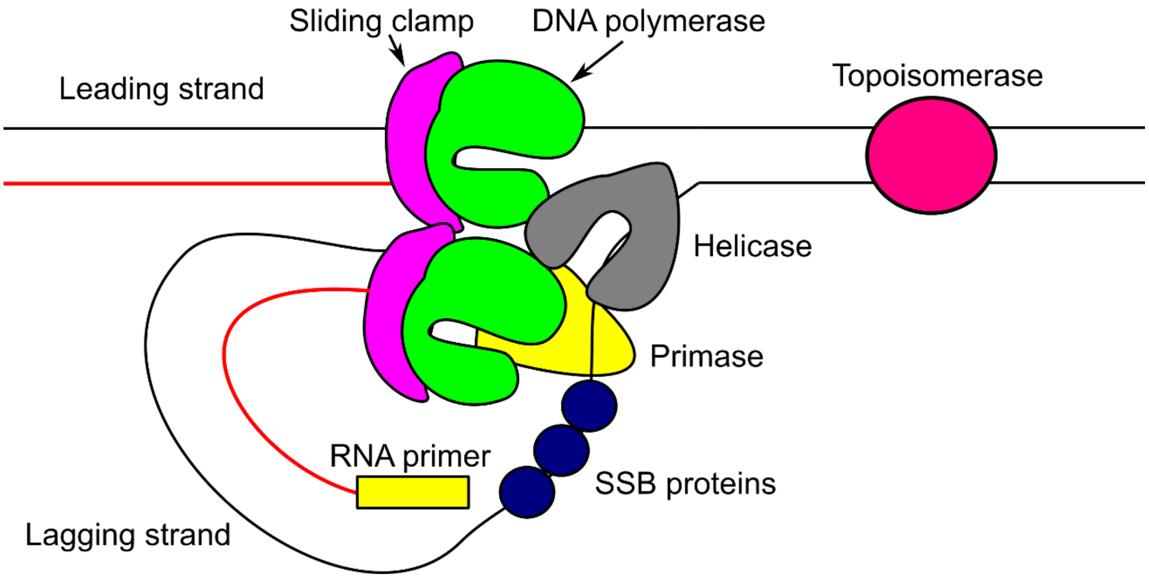

Figure 1.1.1 – Cartoon depiction of DNA replication at the origin of replication. The helicase separates the dsDNA into two complementary strands. Replication proceeds continuously on the leading strand, while the lagging strand is synthesized in short Okazaki fragments.

1.2 Glycosylases and Base Excision Repair

The integrity of genomic information is under constant threat by erroneous chemical modifications that occur ~10,000 times per cell per day from endogenous and exogenous sources[7]. Several DNA repair pathways exist to counter this threat and can be found in all domains of life. One prominent example is the base excision repair (BER) pathway which can handle damage occurring to nitrogenous bases of DNA, as well as, other types of DNA damage[8-10]. In general, the steps of BER include (i) removal of the damaged or modified base by a DNA glycosylase, resulting in an abasic or apurinic/apyrimidinic (AP) site, (ii) cleavage of phosphodiester bond, (iii) generation of the 3'-OH and 5'-phosphate needed for DNA synthesis and ligation, (iv) DNA synthesis to replace to excised nucleotide and (v) ligation of the resulting DNA nick.

DNA glycosylases initiate the BER pathway by employing a nucleotide-flipping strategy (also known as base-flipping) to identify damaged or modified bases and remove them through cleavage of the N-glycosidic bond (Figure 1.2.1)[11, 12]. In many cases this strategy facilitates the search for lesions that may not dramatically distort the overall structure of DNA. There are 11 distinct mammalian glycosylases and while some of these act predominantly on a single type of lesion, such as uracil DNA glycosylase (UDG), others remove many different modified substrates. Three glycosylases act on mismatches involving two canonical nucleotides (G:T mismatches), including thymine DNA glycosylase (TDG)[13], methyl binding domain IV (MBD4)[14] and mismatch DNA glycosylase (MIG)[15]. These glycosylases excise only thymine from G:T mispairs in order to protect against lesions arising from deamination of 5mC to thymine. While much is known on the BER pathway and glycosylases in general, questions still remain on how these enzymes are able to detect lesions amongst an enormous backdrop of normal base pairs.

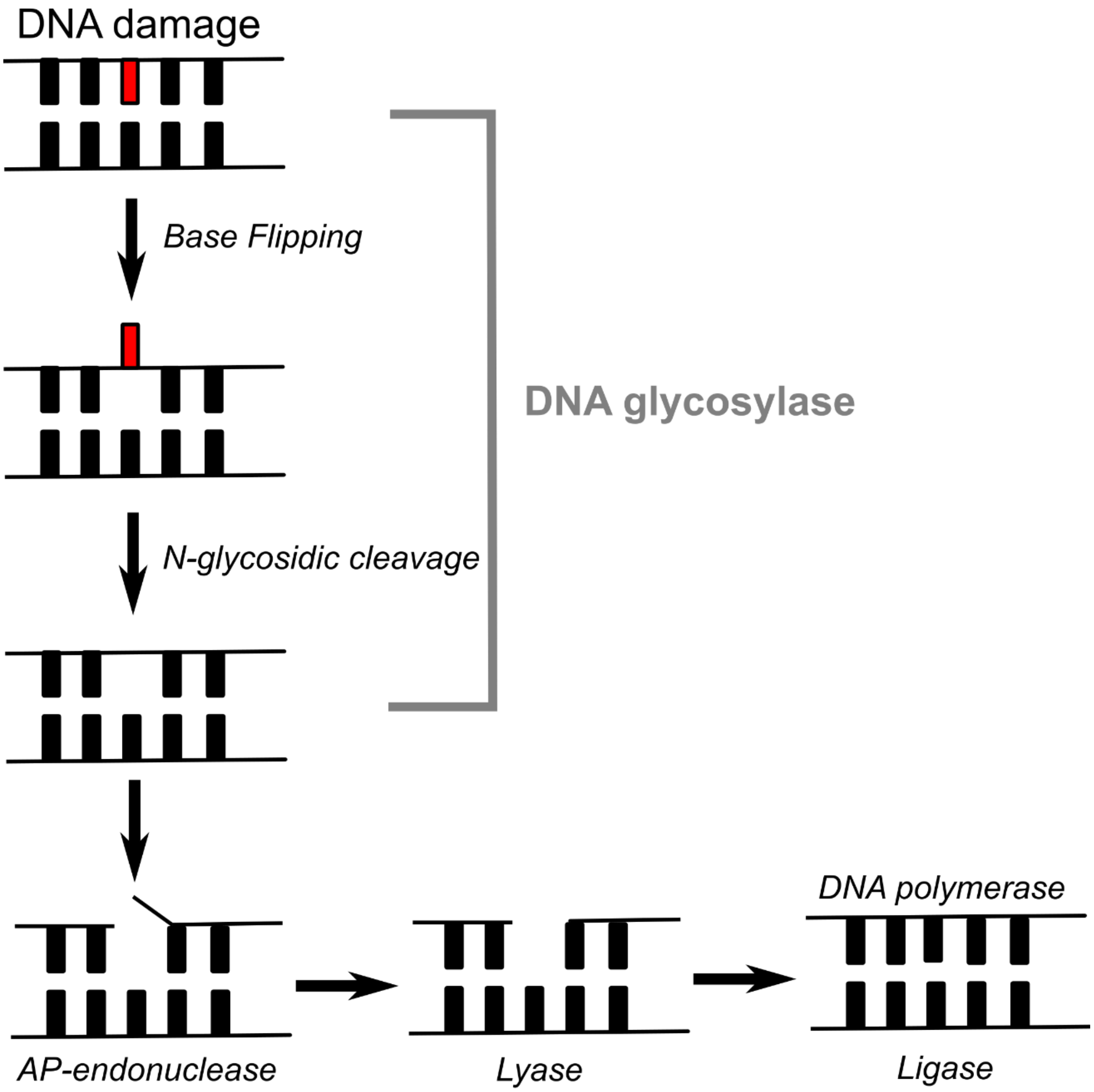

Figure 1.2.1 – Overview of the base excision repair pathway (BER). DNA damage is recognized and removed through DNA glycosylases. This is followed by AP-endonuclease, lyase, DNA polymerase and ligase activity to restore the correct base pair.

1.3 Transcription Initiation

Transcription, the pivotal first step in gene expression, is the process by which a particular segment of DNA is copied into RNA via RNA polymerase. The stretch of DNA to be copied is known as the transcription unit and encodes at least one gene. For proteins, transcription produces messenger RNA (mRNA) to serve as a template for the protein's synthesis through translation. By synthesizing and regulating protein production, transcription plays a fundamental role in everyday cellar activities.

In eukaryotes, the process of transcription begins with RNA polymerase II (Pol II), together with one or more general transcription factors (GTFs) binding to promoter DNA to form the pre-initiation complex (PIC)(Figure 1.3.1). Formation of the closed promoter complex (PIC-CC) is initiated by the transcription factor IID (TFIID) which binds to promoter DNA via contacts between its TATA binding protein (TBP) and the TATA element of the promoter[16]. This is followed by the recruitment of TFIIA, TFIIB, TFIIF and Pol II to the core promoter, ending with TFIIE and TFIIH[17]. Upon formation, the PIC-CC then transitions into the open complex (PIC-OC), in which the melted single-stranded DNA is inserted into the active site and Pol II locates the transcription start site (TSS). The PIC-OC is transient and rapidly converts into the initial transcribing complex (ITC), where mRNA starts to be synthesized[18, 19]. Eventually, Pol II clears the promoter and a stable elongation complex forms, competent for transcription. In recent years, advances in cryogenic electron microscopy (cryo-EM) have led to high-resolution structures of the PIC, thus broadening our understanding of this complex machinery.

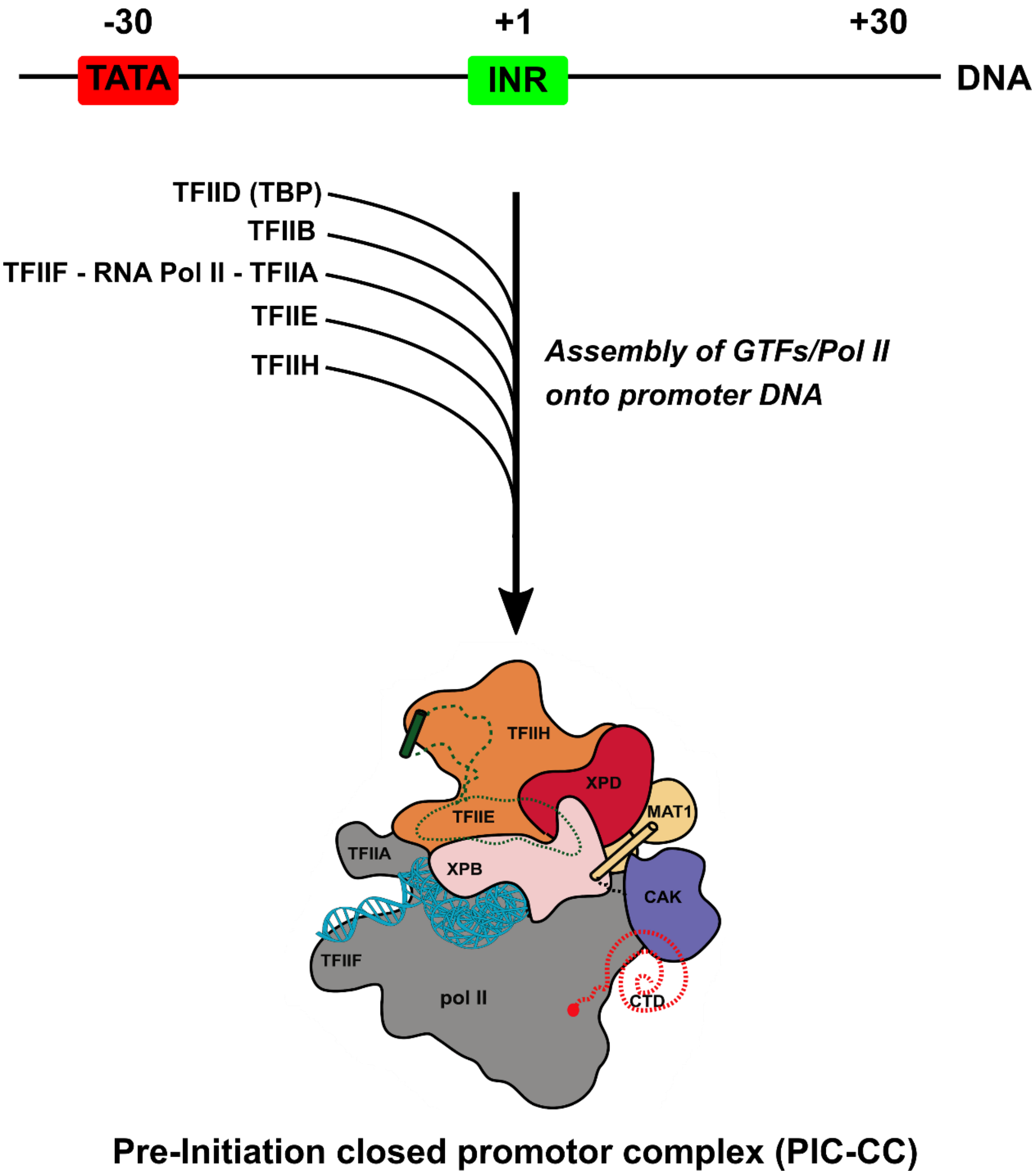
-30
+1
+30
TATA
INR
DNA
TFIID (TBP)
TFIIB
TFIIF - RNA Pol II - TFIIA
TFIIE
TFIIH
Assembly of GTFs/Pol II onto promoter DNA
TFIIH
XPD
MAT1
TFIIE
TFIIA
XPB
CAK
TFIIF
pol II
CTD
Pre-Initiation closed promotor complex (PIC-CC)

Figure 1.3.1 – Schematic depiction of Pre-Initiation complex (PIC) formation. Pol II (gray), along with several transcription factors (labeled and colored) bind to promoter DNA in a step-wise fashion to begin the process of transcription.

1.4 Key Outcomes

The biological systems and processes discussed in this dissertation can be thought of as a smaller subset of the much larger field of genomic duplication and maintenance. In short, the collective theme for this work can be summarized as the study on how the dynamic interactions of these biomolecules influence the conformational changes that govern their everyday functional activities. Notable research accomplishments include:

1. We developed a structural, kinetic and thermodynamic model for the site-to-site transfer mechanism of ssDNA in DNA polymerases during editing. The underlying kinetics and thermodynamics of this transfer mechanism are foundational to maintaining synthesis speeds at the replication fork. As a case study, we investigated DNA polymerase III, the bacterial replicase, which is not only highly accurate but remarkably processive. Moreover, the active sites between the polymerase and exonuclease are separated by ~60 Å. Through path optimization methods, we uncovered the sequence of molecular events that must proceed for translocation of the ssDNA. Employing stochastic modeling techniques, we discovered intermediate (or metastable) states separated by kinetic barriers. We predicted important protein residues using dynamical network analysis. These residues were then tested by biochemical analysis, thereby validating our results. Collectively, this study aids increased our understanding of how polymerases handle errors produced during DNA synthesis.
2. We developed a structural, kinetic and thermodynamic model for the base-interrogation and -flipping mechanisms of thymine DNA glycosylase (TDG). Glycosylases employ unique strategies for recognizing and repairing damaged DNA. Through stochastic

modeling, we showed how TDG employs an intercalating arginine residue to probe the microstructure of dsDNA in search of lesions. TDG then exploits the unstable geometry of the mismatch or lesion through local deformation, allowing the enzyme to effectively "push" the incorrect nucleotide through the minor grove of the dsDNA. Dynamic protein-DNA contacts then facilitate the lesion's ~180° rotation into the enzyme's active site. Collectively, our modeling revealed that this strategy lowers the energetic barrier for base-flipping by ~10 kcal/mol. Moreover, the kinetics of this process occur on a timescale not easily captured by experimental techniques like NMR.

3. Developed a structural model of the human transcription pre-initiation complex (PIC) based on cryo-EM and used the model to make the connection between the PIC functional dynamics and known disease mutations associated with severe genetic disorders. While techniques like cryo-EM and X-ray crystallography have revealed much on the structure of this transcription machinery, these models were incomplete. In this study, we utilized hybrid modeling techniques to build the most complete model of the human PIC to date. This allowed us to characterize undiscovered, yet, important interactions pertinent to the stability of this massive macromolecular complex (22 protein chains). Combining principal component analysis and network analysis then permitted us to partition the dynamics of the PIC into tightly-connected clusters of residues. Importantly, we found that certain genetic mutations appear at critical junctures of dynamic network. Moreover, our results provide a link between the functional dynamics of the PIC and genetic disease.

CHAPTER 2. METHODS

2.1 Molecular Dynamics

2.1.1 The Molecular Dynamics Potential Energy Function

At the core of the MD framework is the potential energy function (or forcefield) which describes the interactions between all atoms in the system of interest. The analytical complexity of atomic interactions results in approximations being made for most quantities (e.g. bonds are treated with Hooke's Law, torsions are treated sinusoidally, van der Waals interactions are computed using an empirically derived 6-12 Lennard-Jones potential). Electrostatic interactions, on the other hand, are represented using Coulomb's Law. Additionally, planarity in aromatic rings can be enforced using improper torsions. Combined, these interactions can be expressed as an N-dimensional function that is dependent on the atomic coordinate space R[20]:

$$V(R) = \sum_{bonds} k_{bond}(r - r_{0,bond})^2 + \sum_{angles} k_{angle}(\theta - \theta_{0,angle})^2 + \sum_{torsions} k_{\phi}[1 - \cos(n\phi - \delta)] + \sum_{impropers} k_{\varphi}(\varphi - \varphi_{0,improper})^2 + \frac{1}{2}\sum_{i}\sum_{j \neq i}\left(\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \varepsilon_{ij}\left(\frac{D_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{D_{ij}}{r_{ij}}\right)^{6}\right) \qquad (2.1)$$

where k represents the force constant, naught-subscripts are equilibrium values, q is the atomic charge, r_{ij} represents the distance between atoms i and j, and D_{ij} is the optimum interatomic distance for van der Waals interactions. The first 4 terms in the forcefield (bonds, angles, torsions and impropers) are referred to as bonded interactions, whereas the last two terms (electrostatic and van der Waals) are referred to as nonbonded interactions. Halving the nonbonded interactions eliminates double counting and ignores self-interactions ($i \neq j$).

A number of different sources allow the parameters of Eq 2.1 to be fitted with some degree of accuracy. Bond and angle force constants are derived from spectroscopic studies. Quantum mechanical (QM) optimizations are employed to compute equilibrium bond lengths, triatomic angles and torsional angles. Torsional angle barriers and Lennard-Jones parameters are fitted to

QM potential energy surface (PES) scans. Finally, atomic charges are typically computed using the RESP method[21].

Over the years a number of different research groups have independently developed forcefields, including AMBER, GROMACS and CHARMM[22-24]. Differences in functionality between these forcefields are almost negligible, thus leaving the choice of forcefield primarily to the user's preference. With the advancements in computer architecture and increasing length of MD simulations, unobserved flaws in the forcefield become apparent and parameters are adjusted accordingly. Additionally, most forcefields are updated frequently to improve fidelity with experimental results.

2.1.2 Force Integration and Trajectory Propagation

MD simulations propagate a collection of atoms through time according to the laws of classical Newtonian mechanics. The acceleration, force and potential energy of a Newtonian system are related to time through,

$$m\frac{d^2x}{dt} = F\big(\boldsymbol{x}(t)\big) = -\nabla V(\boldsymbol{x}(t)). \ (2.2)$$

Where *m* is the mass, ***x*** is the set of atomic positions, *t* is time, *F* denotes the force, $-\nabla V$ is the negative gradient of the potential energy function and the acceleration, ***a***, is equal to $d^2\boldsymbol{x}/_{dt}$. Note that the expressions in Eq. 2.2 are a system of coupled ordinary differential equations (ODEs).

In a purely mathematical sense, this system of ODEs can be readily solved. However, in order to propagate this system through time on a digital machine we need to discretize the time element for some finite number of steps (*n steps*) using a discrete time step (*Δt*, i.e. ***x****(t)* → ***x****(t+Δt)* → ***x****(t+2Δt)* … ***x****(t+nΔt)*). If we choose a small *Δt* then we can expand any function around some arbitrary point and truncate using an expansion series,

$$\boldsymbol{x}(t+\Delta t)=\boldsymbol{x}(t)+\boldsymbol{x}'(t)\Delta t+\frac{1}{2}\boldsymbol{x}''(t)\Delta t^2+\frac{1}{6}\boldsymbol{x}'''(t)\Delta t^3+O\Delta t^4\ ,\ (2.3)$$

$$\boldsymbol{x}(t-\Delta t)=\boldsymbol{x}(t)-\boldsymbol{x}'(t)\Delta t+\frac{1}{2}\boldsymbol{x}''(t)\Delta t^2-\frac{1}{6}\boldsymbol{x}'''(t)\Delta t^3+O\Delta t^4.\ (2.4)$$

Equations 2.3 and 2.4 represent the time reversibility of the system evolution and form the basis for Verlet integration[25], one of the most common propagation algorithms employed in MD simulations. Through a little rearrangement and summation of these two equations, one can obtain,

$$\boldsymbol{x}(t+\Delta t)=2\boldsymbol{x}(t)-\boldsymbol{x}(t-\Delta t)+\boldsymbol{a}(t)\Delta t^2+O\Delta t^4.\ (2.5)$$

Note that the acceleration $\boldsymbol{a}$ enters Eq. 2.5 through the third term in Eqs. 2.3 and 2.4 as the second derivative of the position vector $\boldsymbol{x}$ with respect to time. Conveniently, Eq. 2.5 provides a means to propagate the atomic positions while avoiding the direct calculation of atomic velocities and with an error on the order of Δt^4.

A typical MD simulation begins by collecting the initial positions for all atoms in the form of Cartesian coordinates. Generally, atomic positions are obtained either from X-ray crystallography, NMR or cryogenic electron microscopy (cryo-EM). Additionally, one may obtain initial positions from advanced computational techniques such as homology modeling or *de novo* modeling. To initiate the simulation, velocities drawn from a Maxwellian distribution are assigned to all of the atoms. The forces are calculated through Eq. 2.2. Atoms are then propagated according to Eq. 2.5 one step in time (Δt) and the positions and velocities are updated. The velocities are related to Eq. 2.5 via,

$$\boldsymbol{v}(t)=\frac{x(t+\Delta t)-x(t-\Delta t)}{2\Delta t}.\ (2.6)$$

This process of force calculation followed by propagation is repeated for *n* number of steps which is set by the user.

In order for the Verlet integration formulation to hold, a very small time step needs to be employed. From a MD perspective the time step needs to be such that it captures the fastest molecular motions. Since MD explicitly ignores electrons, the fastest molecular motion corresponds to the vibration of hydrogen bonds, which is on the order of femtoseconds. In most cases, however, the researchers may not be interested in these motions. Thus, these bonds can be constrained around the equilibrium bond length, permitting the use of a 2-fs time step.

As the MD trajectory is propagated through time, instantaneous snapshots of the atomic positions are captured and saved. These represent the time-evolution of the trajectory and can be further analyzed for equilibrium, thermodynamic, kinetic and dynamical properties. Depending on the research problem, snapshots are typically saved between 3 to 6 orders of magnitude longer than the actual time step (2 ps to 2 ns).

2.1.3 Nonbonded Interactions

While bonded interactions can be readily computed, the direct calculation of nonbonded interactions is computationally expensive. The main reason for this discrepancy is that there are potentially O(n^2) nonbonded interactions in a system of *n* particles. For the sake of speed and efficiency, a distance cutoff is enforced for all nonbonded interactions, drastically reducing the computational cost. In the context of van der Waals interactions, the rapid spatial decay of the Lennard-Jones r^{-6} dependence permits the exclusion of all interactions that lie outside the cutoff. Electrostatic interactions, on the other hand, do not decay as rapidly. Thus, these interactions are generally treated with the particle mesh Ewald summation method[26] (PME).

In short, the PME method considers all electrostatic interactions between particles in a unit cell. The Coulombic potential is given by the 5th term in Eq. 2.1,

$$V_{eel} = \frac{1}{2}\sum_i \sum_{j\neq i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}. \ (2.7)$$

Eq. 2.7 can easily be extended to include electrostatic interactions in neighboring periodic cells,

$$V_{eel} = \frac{1}{2}\sum_N \sum_i \sum_{j\neq i} \frac{q_i q_j}{4\pi\varepsilon_0 |r_{ij}+NL|}. \ (2.8)$$

In Eq. 2.8, *N* is the number of periodic cells with a unit cell length of *L* in any given direction. The direct calculation of Eq. 2.8 still presents a computational burden since it converges very slowly. To alleviate this, the Ewald summation separates the distance component *r* into short- and long-range terms,

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1-f(r)}{r}. \text{ (2.9)}$$

Where *f(r)/r* is the short-range term and *1-f(r)/r* is the long-range term. Each particle in the cell is then represented as a point charge with a local neutralizing charge taking the form of a Gaussian distribution,

$$\rho_i(r) = \frac{q_i\alpha^3}{\pi^{3/2}} \exp\left(-\alpha^2 r^2\right). \text{ (2.10)}$$

This results in the short-range term of the Ewald sum taking the form,

$$V_{short} = \frac{1}{2}\sum_N \sum_i \sum_{j \neq i} \frac{q_i q_j}{4\pi\varepsilon_0 |r_{ij}+NL|} \frac{f(\alpha|r_{ij}+LN|)}{|r_{ij}+LN|}. \text{ (2.11)}$$

Neutralizing charges from the Gaussian distribution in Eq. 2.10 are corrected by adding a background distribution, resulting in an expression for the long-range term,

$$V_{long} = \frac{1}{2}\sum_N \sum_i \sum_{j \neq i} \frac{1}{\pi L^3} \frac{q_i q_j}{4\pi\varepsilon_0} \frac{4\pi^2}{k^2} \exp\left(-\frac{k^2}{4\alpha^2}\right) \cos\left(k \cdot r_{ij}\right). \text{ (2.12)}$$

Where *k* are the reciprocal vectors. The free parameter *α* needs to be selected such that it optimizes convergence for both the short- and long-range interactions. Small values of *α* can lead to faster convergence of the long-range interactions, while large values result in the quicker convergence of the short-range interactions.

To improve performance and scaling, PME relies on a fast Fourier transform (FFT) to convert the long-range interactions to reciprocal space resulting in faster convergence with minimal loss in accuracy. Additionally, MD engines like NAMD smooth the reciprocal-space discretized point charges over multiple point grids by employing a Euler spline[27].

2.1.4 Solvent and Simulation Box

Generally, MD simulations can be performed *in vacuo*[28]. However, in the context of biological systems, solvent interactions and electrostatic screening play central roles in the functionality of biomolecules. Currently, there are several computationally efficient water models that exist for MD[29] with the most popular being the TIP3P model. Briefly, TIP3P models the bonds between all atom pairs as rigid bonds set to the equilibrium length. The benefit of this approach is that is removes multiple degrees of freedom while maintaining minimal deviations from experimental bulk properties. This work relies entirely on the TIP3P model for all MD simulations.

Depending on the research problem, MD simulations are typically performed under the canonical (NVT), microcanonical (NVE) or isothermal-isobaric (NPT) ensemble. The N, V, T, P and E are constant variables corresponding to the number of particles, volume, temperature, pressure and energy, respectively. There are multiple schemes that adequately enforce constant temperature and pressure[30-32]. While both the NVT and NVE ensembles possess computational advantages, the NPT ensemble closely replicates experimental conditions. Throughout this work, it is safe to assume that all analysis was performed on trajectory data obtained from MD simulations carried out in the NPT ensemble.

2.2 Enhanced Sampling Methods

While MD is a powerful technique for elucidating structural and dynamic phenomena, the simulation time is limited between the nanosecond and microsecond timescale. Due to its simplicity, MD is not capable of adequately sampling portions of energy landscapes separated by high energy barriers. Moreover, most biological molecules possess complex energy landscapes that contain multiple minima separated by high free energy barriers, leaving the simulation system to become “trapped” in one or more of these minima for long periods of time. The nonergodic nature of the current MD methodology, therefore, does not permit the direct simulation of thermodynamic and kinetic properties for large biological systems with high energy barriers.

In order to address the limitations of conventional MD a number of different methodologies have been proposed[33-36]. These enhanced sampling methods seek to increase the number of rare-event crossings between minima separated by high energy barriers. In most cases, rare-event transitions are accelerated by employing a non-physical force (or bias) which is added to the MD potential energy function. Consequently, this alters the underlying energy landscape in which the statistics at the barriers is significantly increased. Therefore, a *post hoc* reweighting scheme must be applied to enhanced simulation data in order to recover the true thermodynamic and kinetic properties. In the absence of a suitable reweighting protocol, enhanced sampling methods provide an efficient means of phase space exploration.

2.2.1 Accelerated Molecular Dynamics

In 2004, Hamelberg et al. proposed adding a bias potential to the MD potential energy function in order to stimulate infrequent events in molecular simulations[37]. Based on the previous work of Voter[38], accelerated molecular dynamics (aMD) assumes no prior knowledge of the energy landscape. By adding a bias potential to the true potential, aMD modifies the potential energy surfaces near the minima while those near the barrier or saddle point are left unaffected. This, in turn, raises the local minimum surface, thus reducing the dwell time in local minima and increasing the escape rate and chances for phase space exploration. Overall, aMD, provides a robust way to alter the energy landscape while preserving the underlying shape of potential energy surface.

In general, aMD defines a non-negative bias boost potential function, $\Delta V(R)$, such that when the true potential, $V(R)$, is below a chosen threshold (E), the simulation is performed on the modified potential,

$$V^{*}(R) = V(R) + \Delta V(R). \qquad (2.13)$$

Where the modified potential, $V^{*}(R)$, is related to true potential, bias potential and boost energy by,

$$V^{*}(R) = \begin{cases} V(R), & V(R) \geq E \\ V(R) + \Delta V(R), & V(R) < E \end{cases}. \qquad (2.14)$$

Various definitions of the bias potential, $\Delta V(R)$, have been suggested and extensively studied[38, 39]. In the current aMD implementation, the bias potential is defined such that: i) the calculation is computationally inexpensive, ii) the derivative of $V^{*}(R)$ is continuous and iii) the modified potential reproduces the shape of the minima even at a high threshold (E).

$$\Delta V(R) = \frac{(E - V(R))^{2}}{\alpha + (E - V(R))}. \qquad (2.15)$$

In Eq. 2.15, the tuning parameter, α, determines the depth of the modified potential energy basin. Note that when α is zero, the modified potential is flat such that $V^{*}(R) = E$. Additionally, the selection of the tuning and threshold parameters (α and E) determines how aggressively the MD simulation will be accelerated. The current standard is to set the threshold value, E, to a value higher than the average potential energy (V_{min}). The average potential energy, V_{min}, is calculated from a short, normal MD simulation. On the other hand, the tuning parameter, α, is normally set to $E - V_{min}$. This allows the modified potential to mirror the shape of the potential basins.

In this work, aMD is strictly used as a means of phase space exploration for some of the simulation systems. Reweighting of configurations obtained from aMD simulations involving large systems has proved challenging due to large fluctuations in the boost energy. While several reweighting protocols have been investigated, including the use of exponential, Maclauren and cumulant expansions[40], these reweighting schemes rely on small fluctuations in the variance in order to converge to a realistic result. Thus, the recovery of the true canonical ensemble from aMD appears to have only been successful in a handful of cases involving small proteins. Due to this uncertainty, no reweighting of aMD simulation data was performed in this work.

2.2.2 Umbrella Sampling

Another popular method for increasing rare-event transitions in MD is the umbrella sampling (US) method which was first developed by Torrie and Valleau[41, 42]. In US, a bias term is applied to the system along a user-defined reaction coordinate (RC). The RC is defined such that it restricts the sampling to a few degrees of freedom which must adequately describe the transition of interest (i.e. torsions, angles, distances). Additionally, the RC is discretized into multiple segments (windows) that include only a part of the range of the RC. The bias is then applied either to a single simulation or to multiple, independent simulations (windows) in which the distributions overlap.

First, consider that the canonical partition function, Q, of a system can be calculated over the whole phase space via,

$$Q = \int \exp[-\beta V(R)]\, d^N R. \quad (2.16)$$

In Eq. 2.16, $V(R)$ is the potential energy of coordinate system R, β is equal to $1/k_B T$ and N is the number of degrees of freedom. The Helmholtz energy is related to Q by $A = -1/\beta \, ln(Q)$. If a suitable RC can be defined, then the probability distribution along the RC, ξ, can be expressed as

$$Q(\xi) = \frac{\int \delta[\xi(R) - \xi] \exp[(-\beta V)] d^N R}{\int \exp[(-\beta V)] d^N R}. \quad (2.17)$$

Essentially, Eq. 2.17 can be interpreted as the probability of finding the system in a small interval $d\xi$ around ξ. Notably, the free energy along the RC can be readily calculated by $A(\xi) = -1/\beta \, ln(Q(\xi))$, also known as the potential of mean force (PMF). In MD the direct phase-space integrals are impossible to calculate. However, assuming that the system is ergodic, the ensemble average *Q(ξ)* becomes equal to the time average *P(ξ)* for infinite sampling. This leads to,

$$P(\xi) = \lim_{t\to\infty} \frac{1}{t}\int_0^t \rho[\xi(t')]dt' \ . \qquad (2.18)$$

Where *t* denotes the time and *ρ* counts the occurrence of *ξ* in a given interval, typically of finite width.

In US, we consider that *ξ* has been split into multiple windows and that the bias potential, w_i, of some window *i* is an additional energy term that only depends on the RC,

$$V^b(R) = V^u(R) + \ w_i(\xi). \qquad (2.19)$$

In Eq. 2.19, the superscript *u* denotes unbiased, while the superscript *b* denotes biased quantities. Starting with the unbiased distribution, the unbiased free energy, $A_i(\xi)$, can be readily derived to give,

$$A_i(\xi) = \ -\left(\frac{1}{\beta}\right)\ln\left(P_i^b(\xi)\right) - \ w_i(\xi) + F_i. \ (2.20)$$

Where P_i^b is the probability distribution obtained from the biased MD simulation and F_i = -*(1/β)ln(exp[-βw*$_i$*(ξ)])* is independent of *ξ*. The only assumption here is that there is sufficient sampling for each window which is facilitated by the appropriate choice of umbrella potentials, $w_i(\xi)$.

Ideally, the bias potential is chosen such that it ensures uniform sampling within each window spanning the RC. The most common form of the bias term is expressed as a harmonic potential with a force constant of strength *k*,

$$w_i(\xi) = \frac{k}{2}(\xi - \ \xi_i^0)^2. \ (2.21)$$

Here, ξ_i^0, is a reference point which is typically chosen to be the center of each umbrella window. The choice of bias strength (*k*) is a critical decision, one that has to be made prior to the simulation. Overall, the choice of *k* has to be large enough to drive the system over the barrier, yet not so large that it leads to narrow distributions. Overlap between adjacent window distributions is required for some reweighting protocols like the weighted histogram analysis method[43, 44] (WHAM; described below).

In order to recover the unbiased free energy from multiple windows F_i in Eq. 2.20 must be estimated. A widely used method to accomplish this is the weighted histogram analysis method (WHAM). In short, WHAM aims to minimize the statistical error of $P^u(\xi)$ where the global distribution is calculated by a weighted average of distributions of the individual distributions,

$$P^u(\xi) = \sum_i^{windows} p_i(\xi) P_i^u(\xi). \text{ (2.22)}$$

The weights, p_i, are chosen to minimize the statistical error of P^u which leads to,

$$p_i = \frac{a_i}{\sum_j a_j}, a(\xi) = N_i \exp[-\beta w_i(\xi) + \beta F_i]. \text{ (2.23)}$$

Where N is the number of steps sampled for window i. F_i is then readily computed via,

$$\exp(-\beta F_i) = \int P^u(\xi) \exp[-\beta w_i(\xi)]\, d\xi. \text{ (2.24)}$$

In WHAM, the global PMF is obtained by iterating equations 2.23 and 2.24 until convergence. As discussed earlier, the most important requirement for WHAM is sufficient overlap between adjacent window distributions. Failure to meet this requirement will result either in discontinuities or overestimated barrier heights in the combined PMF. While there are several computer programs that perform this calculation, in this work, we make use of the C code distributed by Alan Grossfield[45].

2.3 Path Optimization Methods

Another class of enhanced sampling methods aims to discover the optimal transition pathway or minimum energy path (MEP) between two stable states. Over the years numerous methods have been developed with the purpose of finding suitable transition pathways[46-48]. Generally, these can be divided into two subclasses, in which one relies on the definition of a suitable RC and the other which does not require *a priori* knowledge of the reaction or transition of interest. In this work, we utilized methodologies from both subclasses that represent the path as a collection of configurations (or images) connected by a band or string, namely, the finite temperature string method with swarms of trajectories method (STSM) and the partial nudged elastic band method (PNEB).

2.3.1 Finite Temperature String Method with Swarms of Trajectories Method

In 2007, Roux and co-workers developed the finite temperature string method with swarms of trajectories method (STSM) for discovering minimum energy paths (MEPs) between two stable states[49]. Based on the earlier work of Maragliano et al.[50], STSM builds paths onto the energy surface of the subspace corresponding to a large but finite set of coordinates, referred to as collective variables (CVs). Note that CVs are analogous to what we have defined earlier as the reaction coordinate or RC. Limiting the search to a few degrees of freedom is advantageous since many of the stiff degrees of freedom can be integrated out. If all of the relevant coordinates are included in the definition of CVs, then the MEP becomes an isocommittor path. Moreover, the transition path found constitutes a well-ordered set of states representing the progress of the transition from one basin to another.

In the original formulation, the string is a parameterized curve representing the path and is evolved as a collection of images by estimating the mean force and the metric tensor at each image using constrained dynamics. Notably, STSM follows a very similar formulation. However, in STSM, the string evolves by using a swarm of trajectories initiated from each image to estimate the average drift of each image in CV space.

First, we consider that the probability distribution, Q, along some defined set of CVs, ξ, is given by Eq. 2.6 and that the mean force is related to this probability distribution by $A(\xi)$ = -

1/βln(Q(ξ)). Over some time step, δt, the CVs evolve according to non-inertial Brownian dynamics on a free energy surface,

$$\xi_i(\delta t) = \xi_i(0) + \sum_j \left(\beta D_{ij}[\xi(0)] A_j[\xi(0)] + \delta_{\xi j} D_{ij}[\xi(0)]\right)\delta t + Z_i(0). \text{ (2.25)}$$

Where D_{ij} is the diffusion tensor and $Z_i(0)$ is a Gaussian thermal noise with $\langle Z_i(0) \rangle = 0$. Note that the underlying dynamics that govern the evolution of the full Cartesian coordinates need not evolve with Brownian or Langevin dynamics, but, may instead evolve by Newtonian dynamics as with MD simulations.

We now consider a path, $\xi(\alpha)$, connecting two stable states in a system. The path contains a list of CVs parameterized by the variable α, where $\alpha = 0$ corresponds to the initial state and $\alpha = 1$ is the final state. Accordingly, the system evolves by,

$$\xi_i(\alpha) = \xi_i(\alpha') + \sum_j \left(\beta D_{ij}[\xi(0)] A_j[\xi(0)] + \delta_{\xi j} D_{ij}[\xi(0)]\right)\delta t. \text{ (2.26)}$$

Typically, the path is presented by an ordered sequence of M discrete images, $\{\xi^1, \xi^2, \ldots, \xi^M\}$. Note that if Eq. 2.26 is employed for repeated propagation, the images will also move downhill towards favorable energy regions. The pooling of images in stable basins is undesirable since the goal is to find a MEP that connects two stable states via a high energy transition state. In order to avoid this, a constraint is imposed on each image such that the distance between each neighboring image is equal. This constraint is reinforced at the end of each iteration and allows the reaction path to remain well resolved, especially in high energy transition regions.

Finally, an approximation is needed for Eq. 2.26 to evolve an initial path towards a MEP. The simplest way to accomplish this is to compute the average drift from an ensemble of unbiased trajectories of length δt initiated from each image,

$$\overline{\Delta\xi_\iota(\delta t)} = \sum_j (\beta D_{ij}[\xi(0)] A_j[\xi(0)] + \delta_{\xi j} D_{ij}[\xi(0)])\delta t. \text{ (2.27)}$$

In Eq. 2.27, it is assumed that the thermal noise, $Z_i(0)$, from Eq. 2.25 cancels out. Using this formulation, the system is first thermalized with a bias restraint to keep the CVs near a reference point, ξ^M. The restraints are then released to generate the unbiased trajectory. In this way, the average drift, $\Delta\xi(\delta t)$, can be estimated without making any assumptions about the underlying dynamics of the full Cartesian coordinates. After each evolution, the path is then re-parameterized to satisfy equidistance constraints between images, as with the original string method formulation.

In practice, one iteration of STSM consists of five steps:

(i) Prepare a configuration for each of the *M* images whose corresponding CVs are close to the value ξ^M.

(ii) Generate an equilibrium trajectory for each of the *M* images with ξ restrained around ξ^M.

(iii) Use configurations from the restrained trajectory to run large numbers of short unbiased trajectories for each image.

(iv) Calculate the average drift, $\overline{\Delta z^M}$, to determine the position in CV space of each of the *M* images.

(v) Re-parameterize the path to impose equidistant conformity in CV space for all *M* images.

Generally, the above cycle should be repeated until the images no longer move in CV space. Depending on the transition of interest, this can take many iterations. It should be mentioned that while STSM is a mathematically sound algorithm, success of finding a well-represented MEP is largely dependent on the choice of CVs.

2.3.2 Partial Nudged Elastic Band

An alternative method to STSM is the partial-nudged elastic band method (PNEB)[51]. Briefly, PNEB is a variation of the nudged elastic band method[52] (NEB) which built upon the plain elastic band method proposed by Elber and Karplus[53]. In the original formulation, NEB employs multiple simulations of the system connected by springs to map conformational changes along a path,

$$F_i = F_i^{\parallel} + F_i^{\perp}. \quad (2.28)$$

In Eq. 2.28, the force on each image *i* is decoupled to a perpendicular force, $F_i^{\parallel}$, and a parallel force, $F_i^{\perp}$, by a tangent vector,

$$F_i^{\perp} = -\nabla V(P_i) + \left(\left(\nabla V(P_i)\right) \cdot \tau\right)\tau, \quad (2.29)$$

$$F_i^{\parallel} = [(k_{i+1}(P_{i+1} - P_i) - k_i(P_i - P_{i-1})) \cdot \tau]\tau. \quad (2.30)$$

The tangent vector, τ, defines the path between the initial and final configurations at every image along the path. The dot product of τ with the force field, $\nabla V(P_i)$, represents the contribution of the force field for that image along the path. This is subtracted from the true potential of the image to remove any force contribution along the path from the individual potential of the image. Notably, Eq. 2.29 represents each image as it moves in potential energy space normal to the path, hence the term perpendicular force.

The parallel component, $F_i^{\parallel}$, accounts for the virtual springs connecting the images. In Eq. 2.30, k_i is the force constant and *P* is the positional vector of image *i*. As with the perpendicular force, the tangent vector is used to subtract out the spring forces that act normal to the path. In this way, the spring force keeps the images evenly spaced along the path and does not affect the relaxation of each image. Additionally, maintaining even spacing between images keeps the path well resolved in the saddle point regions, as with STSM.

The only difference between PNEB and the original NEB formulation is that PNEB allows the forces to act on a subset of atoms, thereby excluding the solvent from the calculation. This feature dramatically increases the performance of the algorithm since it limits the optimization to only the relevant part of the system. As with all chain of replicas methods, images along the path are simulated independently, thus, permitting the exploitation of massively parallel computing architectures. Perhaps one of the more important distinctions between PNEB and methods like STSM is that PNEB does not require *a priori* knowledge of the reaction. This means that the user avoids the task of determining a suitable set of CVs to describe the transition. In most systems involving many degrees of freedom this can be a daunting, if not, impossible feat. In this respect, PNEB has a tremendous advantage over CV-based optimization methods like STSM.

2.4 Markov State Models

Over the last several decades a combination of experiment and computation have shown that conformational transitions are essential to the function of proteins and nucleic acids. The length and timescales of these transitions span large ranges and involve complex rearrangements between substrates[54]. In addition to complexity, biomolecular kinetics often involve transitions between a multitude of long-lived, or metastable states that exists on a range of different timescales.

While MD has become an increasingly accepted tool to elucidate conformational transitions in biomolecules, its simplistic formulation does not supply the statistical relevance to accurately quantify the transition rates or kinetics (as discussed previously). To overcome this limitation, a common approach is to partition the conformational space into discrete states. From this partitioning, transition rates or probabilities between states can be calculated, either based on rates theories[55] or based on observations from the MD trajectories[56, 57]. These stochastic models are often referred to as transition networks, master equation models or Markov state models (MSMs), where “Markovianity” means the kinetics are modeled by a memoryless jump process between states[58].

In short, MSMs abandon the view of single trajectories and replace it with an ensemble view of the dynamics. This allows all of the statistical properties to be directly computed since the MSM encodes the ensemble dynamics. Since only conditional probabilities between discretized states are needed to construct the model, simulation trajectories only need to be long enough to reach local equilibrium within the discretized state rather than exceed global equilibrium relaxation times which can be orders of magnitude longer. Thus, long timescale processes can be accurately modeled from simulation trajectories that are orders of magnitude shorter. In the following sections, we describe the individual steps to constructing MSMs, which include: (i) reducing the dimensionality of the MD trajectory data, (ii) discretizing the reduced space, (iii) estimating the transition probability matrix from the discrete states, (iv) agglomerating the discrete states into larger macrostates from the transition probability matrix and (v) calculating the transition rates between macrostates using transition path theory.

2.4.1 Dimensionality Reduction

The first step to constructing an MSM from an ensemble of MD trajectories is to reduce the dimensionality of the MD time series data. While, in principle, one could skip this step and proceed to discretize the full conformational space, this would be impractical due to the high dimensionality and the computational burden it would impose. A robust and automated approach to removing unimportant degrees of freedom is to project the dynamics onto a few ordered parameters using principal component analysis[59] (PCA). Briefly, PCA attempts to find linear combinations of the input coordinates that best explain the variance in the data. This option is attractive since PCA can find orthogonal degrees of freedom that account for the largest amount of variance, while removing degrees of freedom that do not account for much variance in the data. However, in the context of MSMs, PCA suffers from a severe setback in that the underlying assumption is that the kinetically slow degrees of freedom correspond to the high variance degrees of freedom, which is not always the case.

This section describes a dimensionality reduction technique, in the context of MSMs, that utilizes a projection-based metric that is motivated by kinetics. The method, referred to as time-lagged independent component analysis[60, 61] (TICA), was first introduced as a solution to the blind source separation problem. Generally, the goal of TICA is to find linear combinations of the input coordinates that maximize the autocorrelation function of that projection. In addition, each linear combination is constrained to be uncorrelated to previous ones. This is done in a series of maximizations, where each step finds a new independent component (IC) that is slowest subject to being uncorrelated to all of the previously found ICs. Overall, TICA has been shown to dramatically improve the quality of the resulting MSM compared to PCA[62], making it an ideal choice for the dimensionality reduction step in the MSM pipeline.

The theoretical framework begins with $\{\boldsymbol{X}_t\}_{t=0}^{N_f-1}$ which is a multidimensional, discrete time-series. Each snapshot, $\mathbf{X}_t$, is a column vector of dimension *d*, corresponding to vectorized representation of the system conformation. In this notation, N_f is the total number of frames in the trajectory and $\mathbf{X}_t$ is a snapshot at some time *t*. To apply TICA to the time-series both the covariance matrix, $\boldsymbol{C}$, and the time-lagged covariance matrix, $\overline{\boldsymbol{C}}$ are needed,

$$\boldsymbol{C} = \langle(\boldsymbol{X}(t) - \langle\boldsymbol{X}(t)\rangle)^T(\boldsymbol{X}(t) - \langle\boldsymbol{X}(t)\rangle)\rangle, (2.31)$$

$$\overline{\boldsymbol{C}} = \langle(\boldsymbol{X}(t) - \langle\boldsymbol{X}(t)\rangle)^T(\boldsymbol{X}(t - \tau) - \langle\boldsymbol{X}(t)\rangle)\rangle. \text{ (2.32)}$$

In the above equations, angular brackets denote the time averages, the superscript T denotes the transpose and τ denotes the lag time. As with PCA, the generalized eigenvalue problem is then solved,

$$\overline{\boldsymbol{C}}r_i = \boldsymbol{C}\lambda_i r_i. \text{ (2.33)}$$

Where r_i are the independent components (eigenvectors) and λ_i are their respective normalized time autocorrelations (eigenvalues). Note that the time-lagged covariance matrix is usually asymmetric and will generally produce eigenvectors and eigenvalues expressed as complex numbers. In order to avoid this, the time-lagged covariance matrix is symmetrized using ½($\overline{\boldsymbol{C}} + \overline{\boldsymbol{C}^T}$). This is justified by the assumption of time reversibility of the MD trajectory. While it will not be proven here, Eq. 2.33 reveals two important properties of the projected trajectories: (i) the covariance matrix, $\boldsymbol{C}$, is equal to the unity matrix and (ii) the time-lagged covariance matrix, $\overline{\boldsymbol{C}}$, is identical to the eigenvalue matrix λ. The first property conveys that the projected trajectories are normalized so as to have unit variance and that there exists no correlation between any two of them. The second property means that the autocorrelation function of some projection, $a_i(t)$, has a value of λ_i at $\Delta t = \tau$, where Δt is the timestep. Additionally, the cross-correlation function between some projection, $a_i(t)$ and some other projection, $a_j(t)$ $(i \neq j)$, vanishes at $\Delta t = \tau$ and $\tau = 0$. Thus, the eigenvalues can used as a rough estimate to transition timescales through,

$$t_i = -\frac{\tau}{ln|\lambda_i|}. \text{ (2.34)}$$

Clearly, the lag time (τ) is an important parameter, one that must be chosen with great care. Theoretically, the ideal lag time is selected such that it encapsulates degrees of freedom that remain correlated for long timescales while ignoring degrees of freedom that quickly decorrelate. The process of lag time selection is not entirely intuitive and depends primarily on the input coordinates selected for dimensionality reduction with TICA. While small lag times, in most cases, will clearly resolve multiple states in the projected space, they can also give the impression of discontinuities in the sampling which may or may not be the case in the full conformational space. On the other hand, large lag times will begin to incorporate irrelevant degrees of freedom, thereby merging multiple states and masking important kinetic barriers. In practice, a trial and error approach can be employed in which the dimensionality of the input coordinates is reduced systematically using different lag times. This allows the user to assess the effect of lag time on the projected space. Ideally, the optimal lag time will be one that clearly resolves multiple states in the projected space while preserving the kinetic barriers that separate them.

2.4.2 Discretization

MD simulations in the full conformational space are Markovian by construction. However, in practice, the full conformational space is typically projected onto a few ordered parameters (projected or reduced space). The next step is to partition the reduced space into discrete states in order to obtain a computationally tractable description of the dynamics. MSMs then combine these discrete states with the transition probability matrix (discussed in the next section) to model the jump process of the observed trajectory projected onto the discrete states.

Consider a discretization of some state space $\mathbf{\Omega}$ into *n* sets. For practical reasons, the discussion is limited to a simple partition with sharp boundaries. In order to quantify the probability that some point **x** belongs to set *i*, membership functions ($\chi_i(\boldsymbol{x})$) can be defined with the property $\sum_{i=1}^{n} \chi_i(\boldsymbol{x}) = 1$. In the simple partition example, the membership function can be expressed as a step function,

$$\chi_i(\boldsymbol{x}) = \begin{cases} 1, if\ \boldsymbol{x}\ \in\ S_i \\ 0, if\ \boldsymbol{x}\ \notin\ S_i \end{cases}. \quad (2.35)$$

Here, there are n sets $S = \{S_1, \ldots, S_n\}$ which entirely partition the state space ($\bigcup_{i=1}^{n} S_i \in \boldsymbol{\Omega}$) and have no overlap ($S_i \cap S_j = 0$ for all $i \neq j$). An example of this type of partitioning is a Voronoi tessellation, where one defines n centers $\bar{\boldsymbol{x}}_{\boldsymbol{\iota}}$, $i = 1 \ldots n$, and set S_i is the union of all points $\mathbf{x} \in \Omega$ which are closer to $\bar{\boldsymbol{x}}_{\boldsymbol{\iota}}$ than any other center. The closeness of all points to every center is related through a distance metric (e.g. Euclidean distance). The stationary probability of any given point to be in set i is then expressed in terms of the full stationary density,

$$\pi_i = \int \mu\,(\boldsymbol{x}) d\boldsymbol{x}, \text{ (2.36)}$$

where the local stationary density ($\mu_i(\boldsymbol{x})$) restricted to set i is given by,

$$\mu_i(\boldsymbol{x}) = \begin{cases} \frac{\mu(x)}{\pi_i}, \boldsymbol{x} \in S_i \\ 0, \ \boldsymbol{x} \notin S_i \end{cases}. \text{ (2.37)}$$

Notably, these properties are local and thus do not require information about the full conformational space.

In general, there are many clustering algorithms that utilize a Voronoi tessellation and adhere to the discretization formulation described above. These include k-means, k-medoids, regular space clustering, regular time clustering and uniform time clustering. In this work, we make no argument in favor or against any particular clustering method. Any metric that can finely discretize the reduced space as the number of clusters is increased, theoretically, should work. More importantly, a metric should be selected such that it resolves the molecular events of interest, as with dimensionality reduction.

2.4.3 The Transition Probability Matrix

At the core of the MSM framework is the transition probability matrix (TPM). The TPM is a row stochastic matrix that, when combined with the discretization, defines the Markov model. Specifically, the TPM is the discrete approximation of the transfer operator, $\mathcal{T}$,

$$T_{ij}(\tau) = \frac{\langle \chi_j, (\mathcal{T}(\tau) \circ \chi_i) \rangle \mu}{\langle \chi_i, \chi_i \rangle \mu} \text{ . (2.38)}$$

In Eq. 2.38, each element (T_{ij}) represents the time stationary probability to find the system in state j at time $t + \tau$, where τ is the lag time, given that it was in state i at time t. The conditional probability, by definition, is then,

$$T_{ij}(\tau) = \mathbb{P}\left[\boldsymbol{x}(t + \tau) \in S_j \,\middle|\, \boldsymbol{x}(t) \in S_i\right], \text{(2.39)}$$

or,

$$T_{ij}(\tau) = \frac{\int \mu_i(x) p(x, S_j; \tau) dx}{\int \mu_i(x) dx}, for\ all\ x \in S_i. \text{ (2.40)}$$

Notably, Eq. 2.40 reveals that the integrals run over individual sets (S_i) and thus, the only requirement is the local equilibrium distribution ($\mu_i(\boldsymbol{x})$), which is used as weights. This is a very powerful feature of MSMs. Essentially, this means that we do not need any information on the global equilibrium distribution of the system in order to estimate transition probabilities. Additionally, the dynamical information needs only extend over the lag time, τ. Hence, we can estimate the kinetics of a dynamic process from an ensemble of short trajectories, so long as they are at least of length τ and the starting points are drawn from a local equilibrium density. In order to do so, we must first have an expression that describes the change in probability, given that a configuration begins in set S_i and ends in set S_j at some fixed time later ($t + \tau$),

$$\boldsymbol{P} = p_j(t + \tau) = \sum_{i=1}^{n} p_i(t) T_{ij}(\tau), for\ all\ i, j\ \in S. \text{ (2.41)}$$

Note that in Eq. 2.41 the probability of being in set j is obtained through a summation of the probabilities of all configurations in set i multiplied by the conditional probabilities of transitioning from set i to set j. Analogous to the problem encountered in TICA, the resulting matrix, $\boldsymbol{P}$, is symmetrized to enforce detailed balance under the assumption of time reversibility. Moreover, the TPM ($\boldsymbol{P}$) leads to a stationary distribution ($\boldsymbol{\pi}$) by virtue of a simple eigenvalue problem,

$$\boldsymbol{\pi}^T \boldsymbol{P} = \boldsymbol{\pi}^T \text{ . (2.42)}$$

Here, the superscript T denotes the transpose. Eq. 2.42 reveals that the global stationary distribution can be computed from conditional probabilities. Thus, MSMs can correctly recover both equilibrium thermodynamic and kinetic properties of the system, even if trajectories shorter than the longest timescale were used for construction.

As with TICA, the lag time (τ) is a crucial parameter for estimating the TPM, one that will determine the quality and utility of the model. In practice, a suitable value for τ is selected based on the relaxation time of the timescales estimated at different lag times. Short lag times can result in overestimated kinetics and possible discontinuities in the TPM. On the other hand, larger lag times will offset configurations that violate global equilibrium, thus providing higher fidelity but coarser temporal resolution [58].

2.4.4 Simplifying the Transition Probability Matrix

In general, one could immediately proceed to calculating relaxation times (or transition rates) directly from the TPM discussed in the previous the section. However, in practice, the discretization typically produces a large number of states (possibly thousands), resulting in a *n x n* TPM (where *n = number of states*). While the calculation of transition rates from a highly dimensional TPM is computationally feasible, the interpretation of such a kinetic model is impractical. Thus, we would like to be able to simplify the TPM such that it: (i) renders the results more interpretable and visually appealing and (ii) the slow processes are preserved.

A natural approach to simplifying the TPM is to agglomerate the discrete states into larger clusters based on some kinetic metric resulting in a coarse-grained TPM. Eigen-spectrum clustering methods (or spectral clustering) represent an ideal choice since they both simplify the TPM and potentially can preserve information on the slow processes. Initially, Perron-Cluster

Cluster Analysis (PCCA) was proposed for coarse-graining transition matrices[63]. Unfortunately, it was soon discovered that the original PCCA formulation suffered from a severe limitation when it comes to Markov chains in the context of MD. Briefly, PCCA attempts to assign every state in the TPM to a unique cluster based on an eigenvalue decomposition of the TPM. However, transition matrices derived from MD simulations usually contain transition states that cannot be assigned to a unique cluster. Consequently, this results in a large deviation in the TPM from an ideal block structure, a requirement for spectral clustering[64]. Simply put, the corresponding Markov chain is no longer decomposable.

In order to overcome this limitation, the Robust Perron Cluster Analysis (PCCA+) was developed by introducing the concept of fuzzy clustering into the PCCA formulation[65, 66]. In short, fuzzy clustering assigns every object to every cluster with certain probabilities. In contrast to other fuzzy clustering methods, PCCA+ aims to make the clusters as sharp (or crisp) as possible, thus attempting to avoid negative entries in the coarse-grained TPM. A major advantage of PCCA+ is that the coarse-grained TPM obtained from the fuzzy clusters exactly preserves the slow timescales.

The theoretical framework begins with the concept of a simple partition with sharp boundaries, where the membership function ($\chi_i(\boldsymbol{x})$) can be expressed by Eq. 2.35. With fuzzy clustering the condition of discrete values is discarded and the membership function is allowed to take values in the interval [0,1] such that,

$$0 \leq \chi_i(\boldsymbol{x}) \leq 1, \sum_{i=1}^{n} \chi_i(\boldsymbol{x}) = 1 \; . \; (2.43)$$

From here, the data is represented in the form of an undirected similarity graph, $G = (V,E)$, where the vertices ($V = \{v_1, \ldots, v_n\}$) represent the data points. Edges connect the vertices and carry a weight

$w_{ij} \geq 0$, where i and j are two distinct vertices. These values enter the adjacency matrix and the degree (D) of this matrix contains entries along the diagonal,

$$d_i = \sum_{j=1}^{N} w_{ij}. \text{ (2.44)}$$

The weights (w_{ij}) in Eq. 2.44 are obtained from similarities s_{ij}. In practice, there are numerous possibilities to transform the weights to similarities[67]. However, the discussion of these transforms extends beyond the scope of this dissertation.

The next step is to find a partition of the graph (G) such that edges between different clusters have a low weight and edges within clusters have high weight. Note that the number of connected components $A_1, \ldots, A_{nc} \in V$ of some graph G is equal to the multiplicity of the eigenvalue zero of the graph Laplacian,

$$L = I - D^{-1}W. \text{ (2.45)}$$

Where I is the identity matrix, W is the adjacency matrix and D^{-1} is the inverse of the degree matrix. Correspondingly, the eigenspace is spanned by the characteristic vectors $\mathbb{1}_{A1}, \ldots, \mathbb{1}_{AN} \in \{0, 1\}^N$, where $\mathbb{1}$ denotes the vector with all components equal to 1 and,

$$\mathbb{1}_{Aj} = \begin{cases} 1, if\ v_i \in A_j \\ 0, \quad else \end{cases}. \text{ (2.46)}$$

Thus, the most common spectral clustering algorithms proceed through the following steps:

1. Construct a similarity graph with a weighted adjacency matrix W.
2. Calculate a graph Laplacian L.
3. Compute the first n eigenvectors of L.
4. For $i = 1 \ldots N$ let $y_i \in \mathbb{R}^N$ be the ith row of the eigenvectors. Cluster the points $\{y_i\}_{i=1,..,N}$ into clusters $C_1, \ldots, C_N$.

A unique feature of spectral clustering is that it requires only the calculation of the first few eigenvectors. This can be achieved quite readily with standard numerical software, most of which is freely available.

The clustering in step 4 is done with PCCA+, although other choices are possible, including k-means or fuzzy k-means. With PCCA+ the simplex structure of the rows y_i is exploited, making the cluster result independent of any initialization step. Note that the simplex structure would not occur if singular vectors instead of eigenvectors were used for clustering[68].

In the context of MSMs, the TPM is a row-stochastic matrix of the form, $\boldsymbol{P} = D^{-1}W$, and can be interpreted as a transition matrix of a random walk which jumps from vertex to vertex. Moreover, if the graph (G) is connected and non-bipartite then the random walk possesses a unique stationary distribution. Thus, the main idea behind spectral clustering essentially boils down to finding a partition of the graph such that the random walk has long dwell times within the same cluster and seldom jumps between clusters. Additionally, because P and $L = I - P$ have the same eigenvectors, spectral clustering on L is equivalent to clustering on P. The main difference between PCCA+ and other fuzzy clustering methods is that clustering obtained from PCCA+ is the result of a linear transformation of the eigenvectors, which preserves the slow timescales of the random walk[69]. This makes it an excellent choice for coarse-graining TPMs obtained from MD simulations.

2.4.5 Transition Path Theory

The final step in the MSM pipeline is to compute the rates for transitions between states using transition path theory[70, 71] (TPT). For this discussion we will assume that the TPM has been coarse-grained in a manner described previously. To simplify the mathematics, the coarse-grained TPM under consideration contains two metastable states (*A* and *B*) and one intermediate state (*I*), although the theory is generalizable to *n* metastable states.

Prior to computing the rate of the transition from *A* to *B*, we first need to calculate the committor probability, q^+, which is defined as the probability, when being in state *i*, that the system will reach state *B* next rather than *A*. By definition, $q_i^+ = 0$ for all *i* in *A* and $q_i^+ = 1$ for all *i* in *B*. In relationship to the TPM, the committor probability for all intermediate states *i* can be computed by solving the following the system of equations,

$$-q_i^+ + \sum_{k \in I} T_{ij} q_k^+ = -\sum_{k \in B} T_{ik}. \text{ (2.47)}$$

In Eq. 2.47, *T* is the transition probability obtained from the appropriate elements in the TPM. The backward-committor probability, q^-, is then the probability, when being in state *i*, that the system was in state *A* previously, rather than *B*. Under the assumption of equilibrium, this probability is simply $q^- = 1 - q^+$. Notably, the transition probability, T_{ij}, contains contributions from all trajectories. These include trajectories that leave *A* and return to *A* before hitting *B*, or *B* → *A* transitions. Thus, in order to evaluate the statistics strictly in terms of *A* → *B* trajectories, only a fraction of the transitions which come from *A* and go to *B* is relevant (i.e $q_i^- T_{ij} q_j^+$). The probability flux along edge *i*, *j*, contributing to the transition *A* → *B* is then,

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+. \text{ (2.48)}$$

Where f_{ij} is known as the effective flux and π_i is the stationary probability. Unfortunately, the effective flux still contains unnecessary detours such as recrossings. Thus, we need to compute the net flux,

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\}. \text{ (2.49)}$$

In Eq. 2.49, the net flux (f_{ij}^{+}) is a network fluxes leaving states *A* and entering states *B*. Notably, this network is flux-conserving, where the total amount of flux that leaves state *A* will enter state *B* (i.e input flux equals output flux).

From Eq. 2.49 we can compute the total number of observed $A \rightarrow B$ transitions per lag time τ,

$$F = \sum_{i \in A} \sum_{j \notin A} \pi_i T_{ij} q_{ij}^{+}. \text{ (2.50)}$$

Where *F* is the total flux. This is a particularly important quantity since it allows us to directly compute the rate of the $A \rightarrow B$ transition,

$$k_{AB} = {}^{F}\!\big/_{(\tau \sum_{i=1}^{n} \pi_i q_i^{-})}. \text{ (2.51)}$$

An important point to make is that all states that trap the trajectory for some time will reduce the value of k_{AB}. However, these traps are properly accounted for in the total flux, even if they do not contribute to productive pathways. For a more in-depth discussion on the derivation and justification of transition path theory refer to [70, 71].

2.5 Multi-Ensemble Markov Models

One of the limitations of MSMs is that they rely on the underlying MD to reversibly sample rare events. For reasons discussed earlier, high-energy barriers are still out of reach for conventional MD. Enhanced sampling methods like umbrella sampling or accelerated MD, combined with reweighting techniques, can help estimate thermodynamic properties like free energy. However, these reweighting techniques are unsuitable for simulation data with long correlation times in some of the variables, since they treat the input data as uncorrelated samples of the ensemble distribution. To overcome both of these limitations, Wu et al. proposed estimating multi-ensemble Markov models (MEMMs) using the transition-based reweighting analysis method[72] (TRAM). The benefit of MEMMs is that they integrate simulation data from both biased and unbiased ensembles. Moreover, MEMMs estimated with TRAM combine the power of kinetics-based clustering (a feature of traditional MSMs) with the strength of bias sampling to produce thermodynamic and kinetic information at all ensembles.

2.5.1 Transition-based Reweighting Analysis Method

The transition-based reweighting analysis method (TRAM) was developed to integrate simulation data from multiple ensembles in a way that allows it to: (i) work with high-dimensional data and coarse state-space discretizations, (ii) utilize unbiased MD simulations from nonequilibrium starting points and (iii) optimally combine data to full thermodynamics and kinetics at all ensembles. In short, TRAM is a significant improvement to previously proposed transition-based reweighting methods[73, 74], which do not offer all of the above properties. Additionally, methods like WHAM (discussed previously), multistate Bennet acceptance ratio[75] (MBAR), reversable MSMs and discrete TRAM can all be derived from TRAM.

First, consider a molecular system in a reference ensemble with configuration *x* and a dimensionless potential function *u(x)*. The units of *u(x)* are the thermal energy $k_BT = 1/\beta$. Additionally, *u(x)* is a sum of terms, including $\beta V(x)$ and pressure-volume or chemical potential terms. The equilibrium distribution of such a system is,

$$\mu(x) = e^{f-u(x)}. \text{ (2.52)}$$

In Eq. 2.52, the free energy *f* is the negative logarithm of the potential energy function *V(x)*.

Now, suppose that we have simulations from different ensembles. The content of these simulations may comprise an arbitrary combination of unbiased and biased energy functions. Any ensemble can be related to the reference ensemble by introducing a bias potential $b^k(x)$ such that $u^k(x) = u(x) + b^k(x)$. The corresponding equilibrium distribution can then be expressed as,

$$\mu^k(x) = e^{f^k-b^k(x)}\mu(x). \text{ (2.53)}$$

The relative free energy f^k of some ensemble k is chosen such that $\mu^k(x)$ is normalized. In general, the bias potential $b^k(x)$ is selected such that it models commonly used enhanced sampling methods.

We can now expand the earlier discussion on MSMs to include some ensemble k that consists of a partition of the state space into m discrete states $S_1, \ldots S_m$. Additionally, this ensemble contains transition probabilities $T_{ij}^k(\tau)$ that the system is in state S_i at time t will be found in state S_j at time $t+\tau$. The local free energy of state S_i in ensemble k is then,

$$e^{-f_i^k} = e^{f^k} \int \mu^k(x) dx. \text{ (2.54)}$$

Notably, the integral in Eq. 2.54 evaluates to the equilibrium probability of the system to be in state S_i when simulated in ensemble k. Finally, the likelihood of the MSM with transition matrix $\boldsymbol{P}$ is given by,

$$L_{MSM}^k = \prod_{i=1}^m \prod_{j=1}^m (T_{ij}^k)^{c_{ij}^k}. \text{ (2.55)}$$

Where the simulation data from ensemble k contains c_{ij}^k transitions from state S_i at time t to S_j at time $t+\tau$. Recall that one of the assumptions of MSMs is that the underlying dynamics is reversible. This will be the case only for simulations conducted at thermal equilibrium in ensemble k and, thus, will adhere to the detailed balance equations, $e^{f_i^k} T_{ij}^k = e^{f_j^k} T_{ji}^k$. Including detailed balance constraints, the maximum likelihood in Eq. 2.55 has no closed-form solution but can be solved iteratively.

When combining simulation data from multiple ensembles, a central problem is to ascertain the equilibrium distribution at a reference ensemble given the data from all ensembles. The reason behind this is that the equilibrium probability of sample x can be reweighted between different ensembles by means of Eq. 2.53. Methods like MBAR provide optimal estimates of the equilibrium distribution $\mu(x)$ under the assumption that at each ensemble k, the samples x are drawn independently from their global equilibrium distribution $\mu^k(x)$. In contrast to this approach, TRAM does not rely on the global equilibrium assumption, but, instead defines the local equilibrium distribution for each configuration state S_i,

$$\mu_i^k(x) = \begin{cases} e^{f_i^k - f^k}, if\ x \in S_i \\ 0,\ else \end{cases}. \text{ (2.56)}$$

The main assumption in Eq. 2.56 is that the simulations are sampling the local equilibrium distributions. However, these simulations do not need to be in equilibrium with other configuration states, a necessary requirement for employing the MSM framework. The likelihood is then given by,

$$L_{LEQ}^k = \prod_{i=1}^m \prod_{x \in X_i^k} \mu_i^k(x). \text{ (2.57)}$$

Here, X_i^k represents the set of all samples in the *kth* ensemble and in state S_i. Note that μ_i^k can be related to $\mu(x)$ through Eqs. 2.53 and 2.56. Thus, the local equilibrium is key to reweight samples between different ensembles.

The TRAM estimator combines the MSM likelihood and local equilibrium likelihood to give the following,

$$L_{TRAM} = \prod_{k=1}^K (\prod_{i,j} (T_{ij}^k)^{c_{ij}^k})(\prod_{i=1}^m \prod_{x \in X_i^k} \mu(x) e^{f_i^k - b^k(x)}). \text{ (2.58)}$$

Eq. 2.58 represents the probability that a given set of trajectories sampling from different ensembles has visited a particular sequence of discrete states (L^k_{MSM}) and has sampled the local configurations contained within these discrete states. The unknown variables in the TRAM likelihood are $\mu(x)$, f_i^k and T_{ij}^k. The goal of the TRAM estimator is to maximize the likelihood in the unknown variable space subject to three constraints,

$$e^{-f_i^k} T_{ij}^k = e^{-f_j^k} T_{ji}^k. \text{ (2.59)}$$

$$\sum_j T_{ij}^k = 1. \text{ (2.60)}$$

$$\sum_{x \in X} \mu(x) = 1. \text{ (2.61)}$$

Eqs 2.60 and 2.61 are simple normalization constraints. Importantly, the detailed balance condition denoted by Eq. 2.59 couples the MSM part to the local equilibrium part. Thus, the TRAM problem can be thought of expressing two optimization problems simultaneously: (i) optimize the MSMs for given free energies for all configurations at each ensemble k, and (ii) optimize the free energies for all ensembles at each configuration S_i.

In practice, the TRAM problem is transformed into a more tractable system of nonlinear algebraic equations and solved through fixed-point iteration, although Newton-based and other stochastic optimization methods are possible. For a more robust derivation of the TRAM estimator, including the algorithmic details, please refer to [72].

2.6 Dynamical Network Analysis

Dynamical network analysis (network analysis or graph analysis) encompasses a wide array of methodologies based on graph theory. In recent years, network analysis has become a popular tool for examining the dynamics of many body systems, including computer networks, social networks and physical systems[76]. Generally, the individual components of the system under investigation are represented by a node (or vertex). Edges connect interacting components and are assigned a weight indicative of the strength of interaction between the interacting components.

While the definition of nodes in biomolecular systems can be arbitrary, the most common definition is to assign a node to each residue in the protein or nucleic acid. The edges between nodes are determined by contact persistence (typically >90%) and are weighted proportionally to the dynamic correlation between the interacting residues,

$$w_{ij} = -\log\ |c_{ij}|, \text{(2.62)}$$

where the correlation c_{ij} is determined by,

$$c_{ij} = \frac{\sum(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sigma_i \sigma_j}. \text{ (2.63)}$$

In Eq. 2.63, $\bar{x}_i$ and $\bar{x}_j$ denote the time averages obtained from the MD trajectory and σ is the respective standard deviations. Within the constructed network lies key dynamical information on the correlated motions of the biological system. Importantly, these correlated motions can be associated with conformational changes involved in the biological process. Moreover, there exists numerous methods designed to extract pertinent details on the network, some of which will be discussed here.

2.6.1 Community Analysis

One property that many networks have in common is clustering, or network transitivity. This is defined when two nodes that are both neighbors of the same third node have an increased probability of also being neighbors of one another. In terms of biomolecular systems, the network transitivity (or community structure) can be thought of as tightly connected clusters of residues that move together as modules. Traditionally, the community structure of any network can be determined through hierarchical clustering. However, this approach tends to separate single peripheral nodes from the communities to which they rightly belong to.

In order to circumvent the limitation of hierarchical clustering, Girvan and Newman proposed systematically removing edges from the graph based the edge betweenness centrality[77]. The betweenness of an edge is defined as the number of shortest paths between pairs of nodes that run along it,

$$c_B(e) = \sum_{s,t \in V} \frac{\theta(s,t|e)}{\theta(s,t)}. \text{ (2.64)}$$

Where *V* is the set of nodes, *θ(s,t)* is the number of shortest paths connecting nodes *s* and *t*, and *θ(s,t|e)* is the number of those paths passing through edge *e*. The assumption here is that the network contains communities that are loosely connected by a few intercommunity edges. Additionally, all shortest paths between these different communities must go along one of the intercommunity edges. Hence, the intercommunity edges will have a high edge betweenness. By progressively removing these edges, the clusters gradually separate out to reveal the underlying community structure of the graph.

Generally, the Girvan-Newman algorithm proceeds through the following steps:

1. Calculate the betweenness for all edges in the graph.
2. Remove the edge with highest betweenness.
3. Recalculate the betweenness for the remaining edges.
4. Repeat steps 2 and 3 until no edges remain.

More recently, Newman proposed optimizing the modularity[78] as a substitute to the termination protocol (step 4) of removing all edges. Briefly, the modularity of a graph is the fraction of edges that fall within the given communities minus the expected fraction if the edges were distributed at random. Graphs with high modularity have dense connectivity within communities but sparse connections between different communities. Thus, the focus of the algorithm shifts to maximizing the strength of the partition computed at every iteration. In the case of networks derived from MD, a careful balance between optimal modularity and the number of communities found must be observed. This is due to the fact that small changes in modularity can disproportionally increase the number of communities. In the end, striking a balance between the two will produce more interpretable results.

2.6.2 Suboptimal Paths

The transfer of information (or communication) between two nodes spanned by multiple edges can proceed through numerous pathways in a densely connected network. The two nodes under consideration are denoted the source (*s*) and the target (*t*), respectively. Networks derived from MD are undirected and therefore the source and target are interchangeable. Consider now the shortest path between *s* and *t*, which is also the optimal path. In addition, there exists a number of slightly longer, albeit, nearly optimal paths that will contribute to communication between these two nodes (e.g. residues from two distal active sites on a protein). In order to accurately map the flow of information between *s* and *t*, these suboptimal pathways must be taken into account[79]. For MD networks, these paths are responsible for the bulk of allosteric communication between *s* and *t*.

The length *l* of any given path between *s* and *t* is given by the summation of all the edge weights w_e connecting *s* and *t*,

$$l = \sum_{e \in s|t} w_e. \text{ (2.65)}$$

Recall that the edge weights are proportional to the dynamic correlation between any two nodes *i* and *j* (Eq. 2.62). One reason for computing the shortest pathways is that the negative logarithm of strongly correlated residues results in shorter distances. The code used in this work finds all suboptimal paths that are longer than the optimal path, yet, shorter than a user-defined cutoff value. Thus, it is assumed that all pathways that fall within this cutoff contribute significantly to allosteric communication.

CHAPTER 3. POLYMERIZATION AND EDITING MODES OF A HIGH-FIDELITY POLYMERASE ARE LINKED BY A WELL-DEFINED PATH

The work presented in this chapter was performed in collaboration with Meindert H. Lamers, Margherita Botto, Rafael Fernandez-Leiro and Fabian Paul. All computational work was conducted by Thomas Dodd, while the biochemical experiments were performed by Margherita Botto. This chapter of the manuscript was written with input from all persons named above.

3.1 Abstract

Proofreading by replicative DNA polymerases is a fundamental mechanism ensuring DNA replication fidelity. In proofreading, mis-incorporated nucleotides are excised through the 3'-5' exonuclease activity of the DNA polymerase holoenzyme. The exonuclease site is distal from the polymerization site, imposing stringent structural and kinetic requirements for efficient primer strand transfer. Yet, the molecular mechanism of this transfer is not known. Here we employ molecular simulations using recent cryo-EM structures and biochemical analyses to delineate an optimal free energy path connecting the polymerization and exonuclease states of E. coli replicative DNA polymerase Pol III. We identify structures for all intermediates, in which the transitioning primer strand is stabilized by conserved Pol III residues along the fingers, thumb and exonuclease domains. We demonstrate switching kinetics on a tens of milliseconds timescale and unveil a complete pol-to-exo switching mechanism, validated by targeted mutational experiments.

3.2 Introduction

Replicative DNA polymerases synthesize new DNA with extraordinary fidelity[80, 81]. Incorrect nucleotide insertion into the growing primer strand occurs at a rate not exceeding one per 10^6 synthesized bases. Three distinct features of the DNA polymerase holoenzyme are responsible for this remarkable precision[82-85]. First, polymerases' active sites have evolved to select for the nucleotide with correct Watson-Crick base pairing to the template strand. Second, after mismatch incorporation, the growing end of the primer terminus becomes misplaced, preventing further DNA extension. Third, the mismatch presence induces DNA fraying at the primer-template junction[86], promoting release of the primer end from the polymerase active site. Together, this outcome alters the equilibrium between DNA synthesis (polymerization) and excision by the 3'– 5' exonuclease subunit (editing or exonuclease activity).

Removal of mis-incorporated nucleotides is essential for accurate genome duplication. Yet, the molecular mechanism of transferring the primer end from the polymerase to the exonuclease active sites remains elusive. In a recent breakthrough, cryo-EM captured the bacterial DNA polymerase III (Pol III) core in both the polymerase and exonuclease functional states[86, 87], shedding light on the conformational changes that must accompany pol-to-exo mode conformational switching. While informative, the new structures visualize only the end states of the switching transition and, thus, do not explain how the primer end traverses the ~60-Å distance separating the two active sites.

To understand the mechanism of this process vital for genome stability, we focus on the core of the Escherichia coli Pol III holoenzyme, composed of the α, ε and θ subunits (Fig. 3.2.1). Similar to other C-family polymerases, the α subunit[88] holds the polymerization site and has a characteristic shape resembling a right hand with fingers, thumb and palm domains[89-91]. The α subunit also has a Polymerase and Histidinol Phosphatase (PHP) domain. Known to function as the exonuclease in most bacteria, the PHP has been inactivated in proteobacteria such as E. coli[92, 93]. Instead, the ε subunit serves as the 3'–5' exonuclease and is directly attached to α by the thumb and PHP domains[94, 95]. The θ subunit has no enzymatic function but binds and stabilizes ε[96-98]. The Pol III core (α, ε and θ) binds to the DNA sliding clamp β[99, 100], essential for processive DNA synthesis. DNA synthesis by Pol III core – β complex is fast (600-1000 nucleotides per second), processive (>100,000 nucleotides per binding event) and at the same time highly precise (error rate ~ 1 per million) [101, 102, 95, 103].

Modern computational science offers powerful tools to expose the microscopic dynamics underlying complex biomolecular transitions, provided that structures for the initial and final states are known. Specifically, in this study we relied on chain-of-replicas path optimization[51, 104-106] to compute a minimum free energy path connecting the polymerization and proofreading states of the Pol III holoenzyme, in which the DNA construct had a G:T mismatch at the primer end. Applying path optimization methods to large macromolecular complexes was, until recently, computationally prohibitively expensive. Advances in GPU technology and massively parallel computing platforms made it possible to use molecular dynamics (MD) to sample the conformational ensemble along the precomputed path (>6 μs of combined unbiased and biased sampling). We then employed the transition-based reweighting analysis method (TRAM) to construct a multi-ensemble Markov model (MEMM)[52, 107] from the MD trajectory data.

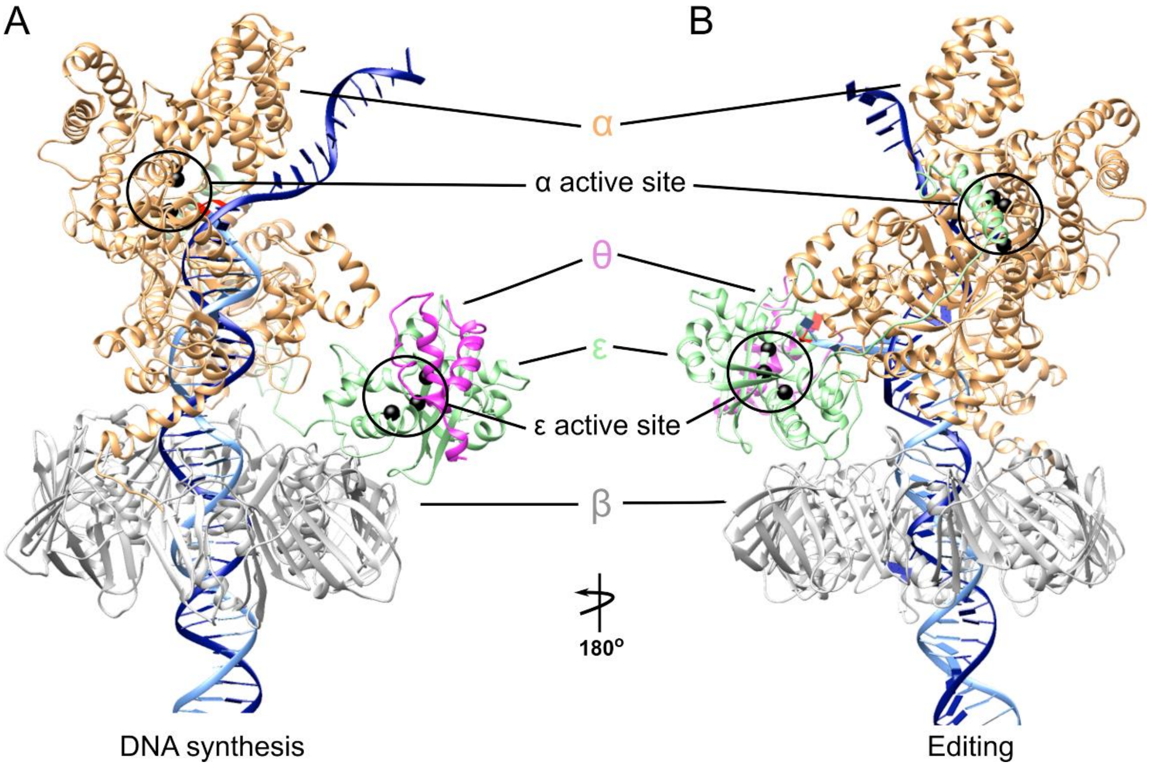

Figure 3.2.1 - Overview of DNA polymerase III in both polymerization (left) and editing (right) modes. Subunits are colored and labeled. Active sites in both the α and ε subunits are highlighted with circles.

The MEMM yields a complete kinetic model for the pol-to-exo mode conformational switching, including transition rates for all on-path intermediates. After partitioning the conformational ensemble into distinct kinetic macrostates, we applied dynamic network analysis to each macrostate. Key residues (critical nodes) along the path of the transitioning primer were determined, extending from the α subunit palm and thumb domain to the ε subunit. To validate the computational models, knowledge of the critical nodes was combined with data from conservation analysis to design mutations that disrupt the ordered transfer of the primer end to the exonuclease site and, thus, affect the balance between DNA synthesis and editing. Collectively, our results unravel the molecular origins of Pol III holoenzyme efficiency and fidelity.

3.3 Results

3.3.1 Pol III holoenzyme transitions from pol to exo mode along a well-defined path

To model the Pol III holoenzyme conformational transition from polymerization to editing, we started with the end point conformations captured by cryo-EM[86, 87]. We built models for the two end states (denoted pol and exo, respectively), comprised of Pol III core, the β-clamp and primer-template DNA with a G-T mismatch at the primer end. We then used molecular dynamics flexible fitting (MDFF) with a weak scaling factor (ξ=0.1) to extensively equilibrate the models, while ensuring conformance to the respective EM densities. A short targeted MD run was used to connect the equilibrated end states. From the targeted MD trajectory we selected 32 evenly spaced snapshots (replicas) that served to initiate our path optimization protocol, employing the partial nudged elastic band method (PNEB)[52, 104]. In PNEB, the minimum energy path connecting protein functional states is represented by a series of replicas of the simulation system. PNEB uses forces from MD to optimize the protein conformations in all replicas to minimize the energy gradient perpendicular to the path. Forces applied parallel to the path keep the conformations in neighboring replicas distinct, while allowing the path to sample favorable regions of the free energy landscape. In this instance, we ran PNEB until convergence with 32 replicas representing the path, accumulating 18 ns of sampling per replica.

Our computed MEP delineates the sequence of molecular events and precise conformational shifts that transition the Pol III core from a pol to exo state. The process begins by fraying of the mismatched G-T pair at the primer terminus. To reach the exonuclease active site, three nucleotides must unpair at the primer-template junction and extend toward the ε subunit. Indeed, apart from G-T mismatch fraying, we observe two additional sequential unpairing events. However, DNA fraying and unpairing is not sufficient to accomplish this transition. In polymerization mode, a 3-nucleotide ssDNA overhang, even if fully extended, would not be able to span the ~70-Å distance to the exonuclease site. Instead, we observe 5.3 Å backtracking and 32.9° rotation of the DNA duplex that occupies the central cavity formed by the ring-shaped β-clamp and the Pol III core (Fig. 3.3.1A and 3.3.1B). Importantly, the Pol III core takes advantage of the spiral motion of dsDNA inside the cavity – a motion which is also essential for successful primer extension during replication. The Pol III core accommodates this motion by presenting positively charged residues along the entire length of the bound dsDNA[87]. These transient contacts track along the DNA backbone and facilitate the forward rotational movement of Pol III during DNA synthesis. Upon encountering a mismatch, a continuation of the spiral motion, but without addition of nucleotides brings the fraying primer terminus in proximity to the exonuclease site. The third element necessary for pol-to-exo mode switching is the conformational change of the Pol III core itself.

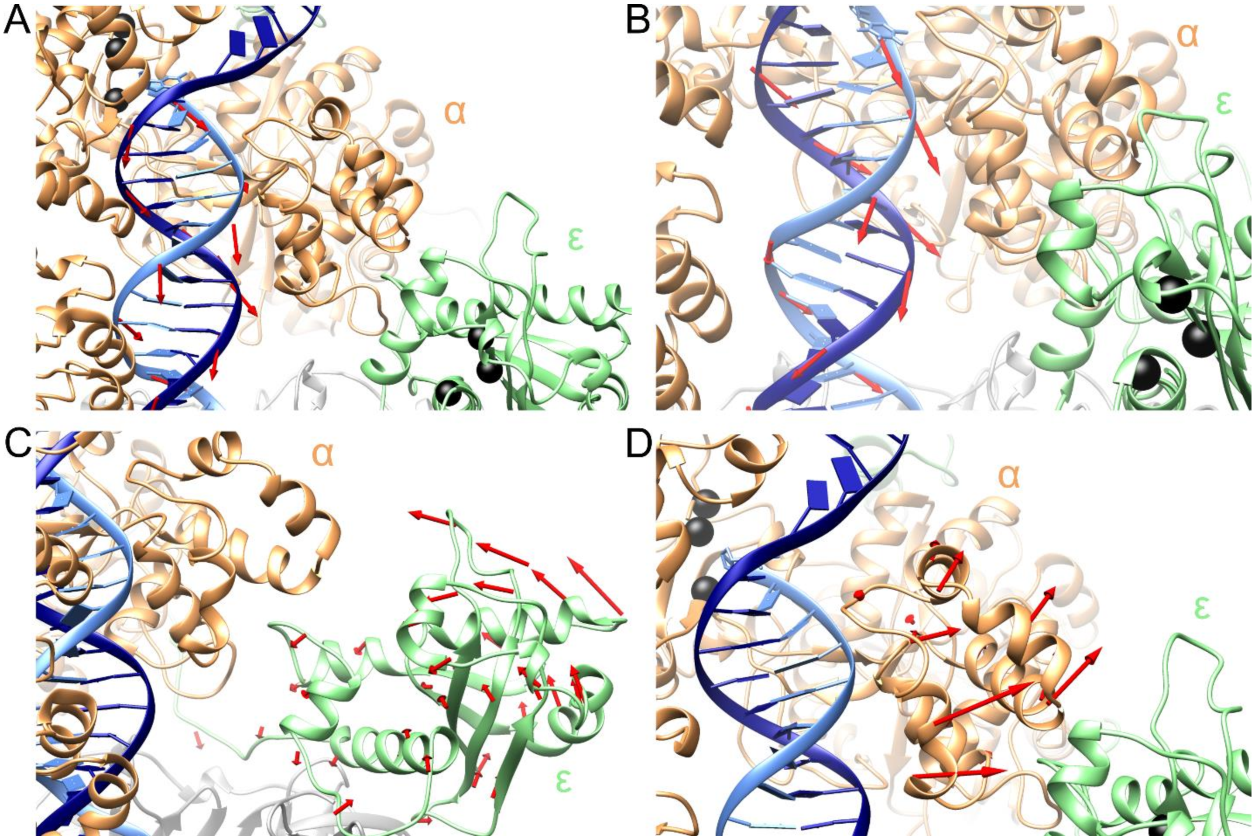

Figure 3.3.1 - Concerted motions of Pol III holoenzyme guide the primer along the path toward the exonuclease state. A, Initial backtracking motion of the DNA duplex away from polymerase active site observed in the MEP. B, Subsequent rotational motion of the DNA and additional backtracking facilitates sequential unpairing at the primer/template junction. C, Tilting motion of the ε subunit toward the α subunit shortens the distance between the polymerase and exonuclease active sites. D, Outward shift of the thumb domain with respect to the PHP domain creates an opening to accommodate repositioning of the ε subunit. Red arrows indicate direction of motions observed in the minimum free energy path (MEP). Shifts in atomic positions for consecutive replicas of the MEP during different stages of the pol-to-exo transition were computed as vectors and mapped onto the structural elements of the Pol III holoenzyme. The α subunit is shown in orange; the ε subunit is shown in light green; the primer and template DNA strands are shown in light and dark blue, respectively; residues in the pol and exo active sites are shown as black spheres. The θ subunit has been omitted for clarity.

Specifically, in the minimum energy path we observe an ~12^o tilting movement of the ε subunit toward the α subunit, which shortens the distance to the exo site by ~10 Å in the pol-to-exo mode transition (Fig. 3.3.1C). Furthermore, the thumb domain moves outwards to make space for the passing primer (Fig. 3.3.1D). While the thumb domain's role as a steric wedge to separate the DNA strands is unique to the C-family of DNA polymerases, its functional significance has been highlighted in both A and B-family DNA polymerases[86, 108, 109]. Here, our computational modeling sheds light on a new role for the thumb domain, which is to create an opening to accommodate the shift of the ε subunit.

3.3.2 Stable intermediates and a complete kinetic model for the pol-to-exo mode transition

PNEB optimization produces a time ordered series of structures, representing the Pol III pol-to-exo mode transition in its entirety. Next, we used these structures as seeds to initiate free molecular dynamics simulations and extensively sample the conformational ensemble along the optimal path. Since unbiased MD yields an ensemble obeying Boltzmann statistics, it becomes possible to analyze this ensemble and identify metastable states, corresponding to stable intermediates along the path. Moreover, the MD trajectories hold information on all observed state-to-state transitions, which allows us to construct kinetic models linking the on-path metastable states.

Prior to estimating kinetic rates from our simulation data, we carried out time-lagged independent component analysis[61] on the unbiased MD trajectories to identify the slowly varying degrees of freedom associated with the pol-to-exo conformational transition. Select atomic distances between the primer strand and the α/ε subunits (Materials and Methods) were computed along the unbiased MD trajectories. Time-lagged independent components (ICs) were obtained from this distance data and all trajectory frames were projected onto the first two ICs (Fig. 3.3.2A).

Under-sampled regions in the space defined by the two ICs indicated the presence of significant energy barriers in the pol-to-exo mode transition. Umbrella sampling was then selectively applied only to these barrier regions of the free energy landscape.

To combine the biased and unbiased MD data, we employed the transition-based reweighting analysis method (TRAM)[72], a recently introduced statistically optimal approach to estimate multi-ensemble Markov models (MEMM)[72, 107] with full thermodynamic and kinetic information at all ensembles. The approach combines the benefits of Markov state models[110, 111, 71, 112] - kinetics-based clustering of high-dimensional data and modeling of complex many-state systems - with the strength of biased MD to accelerate rare event sampling. The method has been shown to yield reliable microstate free energies and accurate kinetic rates on timescales of milliseconds to seconds, directly comparable to experiment[107]. We constructed an MEMM, which partitioned the conformational ensemble into 8 kinetically distinct macrostates (denoted S1-S8, Fig. 3.3.2B). We then computed probability fluxes and estimated transition timescales in and out of each macrostate. The end result was a complete kinetic model for pol-to-exo conformational switching (Fig. 3.3.3). Notably, we found that primer translocation to the exonuclease site occurs on an overall timescale of ten milliseconds, exceeding the timescale of nucleotide incorporation by an order of magnitude. Thus, Pol III core achieves a delicate balance: the rate of conformational switching is slow enough not to interfere with normal nucleotide incorporation, and yet minor stalling upon mismatch encounter causes efficient transfer and removal of the incorrect nucleotide by the Pol III ε subunit. The effective free energy landscape along the two ICs (Fig. 3.3.2A) indicates a stepwise pol-to-exo mode transition with clearly resolved DNA melting and primer end translocation events. The process starts from state S1 (polymerization mode), proceeding through two early intermediates S2 and S3, in which the terminal G-T base pair is unraveled.

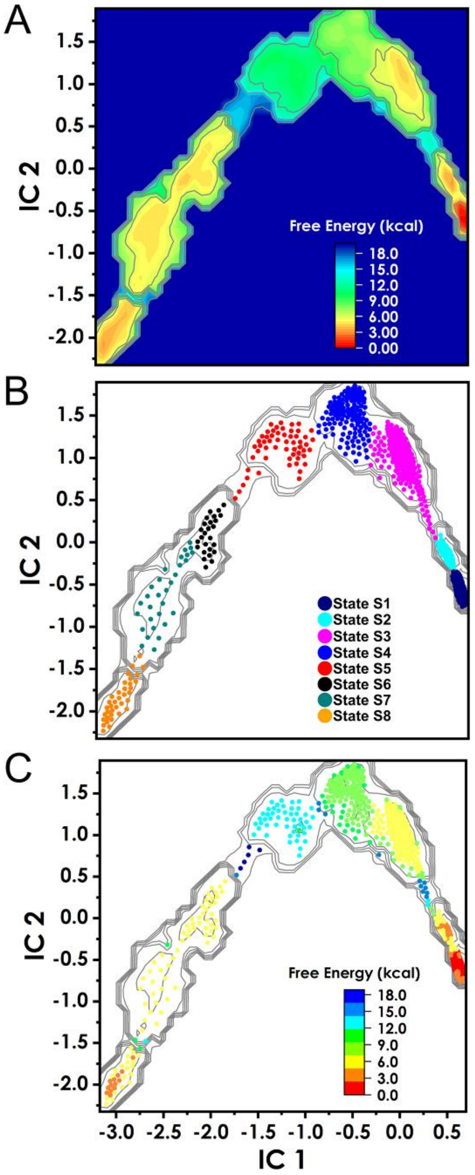

Figure 3.3.2 - Analysis of the Pol III conformational ensemble reveals distinct kinetic 568 intermediates in the pol-to-exo transition. A, Effective free energy profile projected onto the first two independent components (ICs) from TICA analysis. Inset denotes ΔG scale in kcal/mol and is set relative to the polymerization state. B, Multi-ensemble Markov model (MEMM) constructed by combining the biased and unbiased simulation ensembles. Microstates (dots) are colored by the macrostate (intermediate) they belong to. Macrostate identities were computed with the PCCA+ algorithm. Color scheme for the macrostates is shown in the inset. C, Microstates (dots) colored by their computed free energies from the MEMM analysis. Inset denotes ΔG scale in kcal/mol and is set relative to polymerization state.

While in state S2 the frayed primer end is still proximal to the polymerization site, in state S3 the mispaired T base has rotated away by 6 Å, effectively preventing DNA synthesis. Additional DNA translocation along the DNA axis by ~7 Å in S4 leads to an intermediate with completely open G-T base pair and partially disrupted hydrogen bonding for the second base pair from the primer end. DNA backtracking and rotation are facilitated by a patch of positively charged residues from the extended fingers domain (K839, R876, R877 and K881) that make contacts with the downstream DNA duplex. The highest barriers in the free energy landscape correspond to unpairing of the second and third nucleotide from the primer end (S4-S5 and S5-S6 transitions). The respective saddle point regions are 10.9 and 15.5 kcal/mol higher than the initial state S1 (Fig. 3.3.2C), resulting in the slowest computed timescales of 2,100 μs and 5,100 μs. Starting in S4, residues from the Pol III thumb domain insert between the template and the primer end serving as a wedge to separate the two strands. In states S5 and S6, a positively charged patch on the surface of the thumb domain (K439, R443, R447, K461) binds and provides electrostatic stabilization for the transitioning DNA primer overhang. The final stages of primer translocation (S6-S7 and S7-S8 transitions), involve tilting of the Pol III thumb domain away from the dsDNA and the π-stacking of a tyrosine (Y453) onto the last base pair of the DNA duplex. Together, these conformational shifts induce strain in the downstream DNA duplex and further increase the separation of the primer and template strands. In state S7, the terminal thymine base contacts a hydrophobic residue cluster from the ε subunit (M18, V65, F102). Between states S7 and S8, we observe closing of the gap between the α and ε subunits, allowing the primer end to insert into the exonuclease active site in a catalytically competent orientation. The timescale for this transition is comparatively slow (711.1 μs and a free energy barrier of 8.2 kcal/mol), suggesting that primer insertion is gated by the motion of ε subunit within the Pol III core. Indeed, in previous studies the ε subunit was found to be relatively mobile during pol-to-exo switching due to the weak interactions of ε with the β clamp[113, 114].

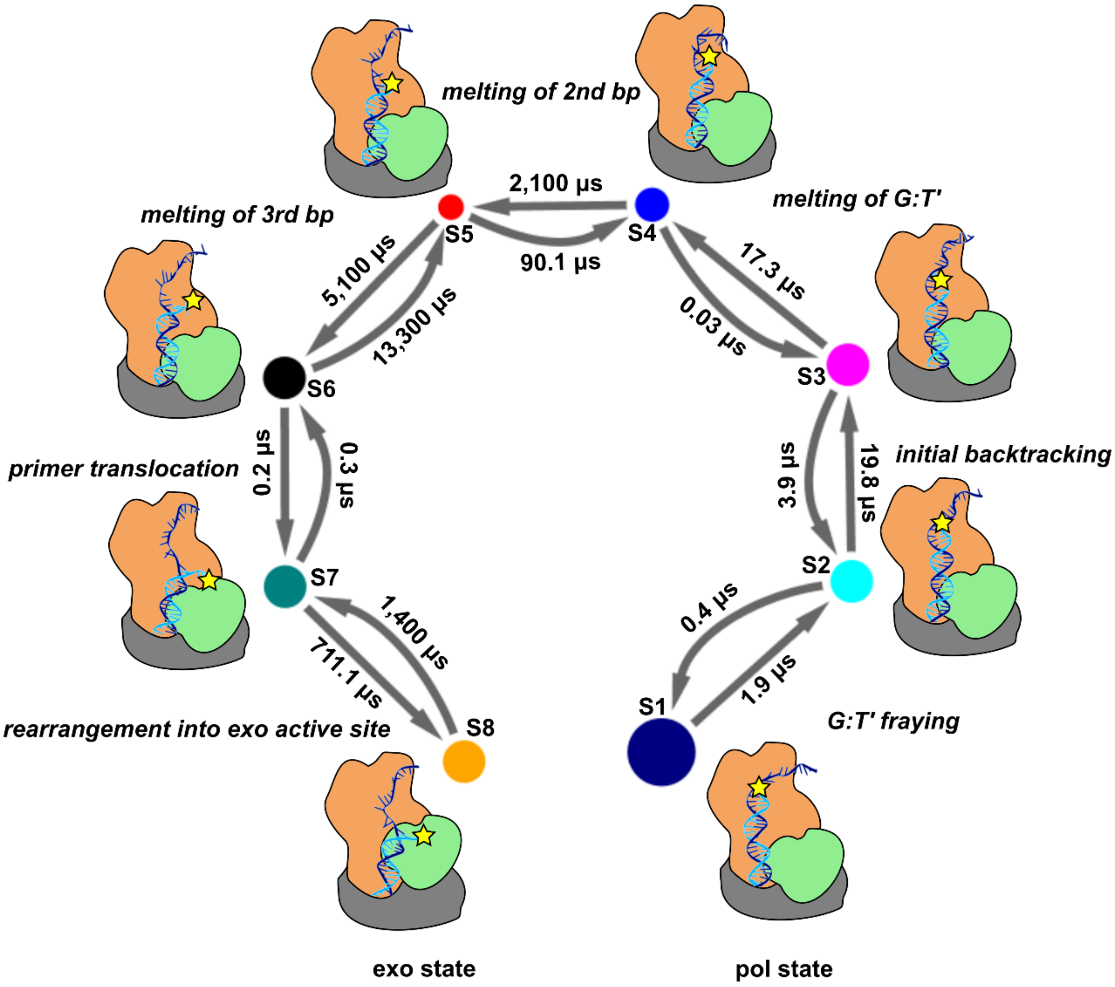

Figure 3.3.3 - Complete kinetic model for the pol-to-exo mode transition connecting all on-path intermediates identified by the MEMM analysis. Macrostates S1-S8 are denoted by circles. Larger circles correspond to more populated macrostates. Transition between states are indicated with arrows and computed timescales for transitioning in and out of each macrostate are shown above the arrows. Each microstate is also represented by a cartoon, indicating the position and the extent of unpairing of the DNA primer end. The position of the mismatch nucleotide on the primer strand is indicated by a yellow star.

When the primer strand is bound to the exonuclease active site, most contacts are to the terminal nucleotide, with an important hydrogen bond between the 3' hydroxyl of the ribose and the backbone nitrogen of Threonine 15[115]. Therefore, once the bond between the terminal and adjacent nucleotide is cleaved, there are few interactions that keep the primer strand within the exonuclease. With the mispaired nucleotide removed and only two melted nucleotides remaining, the return to the polymerase active site will be swift, enabling DNA synthesis to resume without delay.

3.3.3 Critical residues in the pol-to-exo conformational transition

The MEMM results allowed us to analyze each kinetically distinct macrostate and dissect the precise interactions, dynamic rearrangements and residue networks underlying the switching mechanism. Knowledge of the detailed mechanism served as a basis for successful validation of our computational models. Specifically, we employed dynamic network analysis to partition the holoenzyme complex into dynamic communities (tightly connected clusters of residues that move together as modules), mapping protein and nucleic acid residues onto graphs wherein each residue is a node and contacting nodes are connected by edges. All edges are weighted by dynamic correlation. Using these graphs, we computed suboptimal paths[116, 117] connecting the polymerization and exonuclease active sites for states S1-S8. Suboptimal paths are a set of paths with length shorter than a specified limit above the optimal path. Suboptimal paths reflect residue correlations in molecular dynamics and, thereby, offer a way to quantify allosteric communication. Furthermore, nodes traversed by the largest number of suboptimal paths frequently correspond to critical residues for allosteric communication and regulation. Critical residues in the Pol III core identified by the above analysis were also tested for amino acid conservation and persistent contacts between DNA and the α or ε subunits. Combined residue scores were obtained from the individual suboptimal path score, conservation score and contact persistence score. Highest scoring residues were then mapped onto the Pol III holoenzyme structure (Fig. 3.3.4A and Materials and Methods). Critical residues were found in the palm and thumb domains, for which we posit multiple roles in pol-to-exo conformational switching (Fig. 3.3.4B-D). In pol mode, residues R443, R447 and K510 form contacts to the DNA minor groove, while also stabilizing the separated template and primer strands during the latter stages of the transition. Pol III thumb domain residues Y453 and K461 stabilize the separated primer. The importance of Y453 was noted in previous experimental studies[86, 118]. We also noted a loop in the thumb domain (P464-M469) that protrudes into the DNA major groove in polymerization mode, while directly binding the template strand in exonuclease mode. We posit that the P464-M469 loop restricts the movement of the DNA duplex during replication while during the pol-to-exo transition it serves to anchor the template strand, ensuring strand separation prior to exonuclease excision. Palm domain residues R411, H511 and R560 may serve similar roles, interacting with the DNA minor groove and the template strand. We also identified a hydrophobic cluster at the opening of the ε active site (M18, V65, F102; Fig. 3.3.3.1E) that transiently stabilizes the primer end prior to insertion into the exonuclease site – a process which is dynamically gated by the motion of the ε subunit.

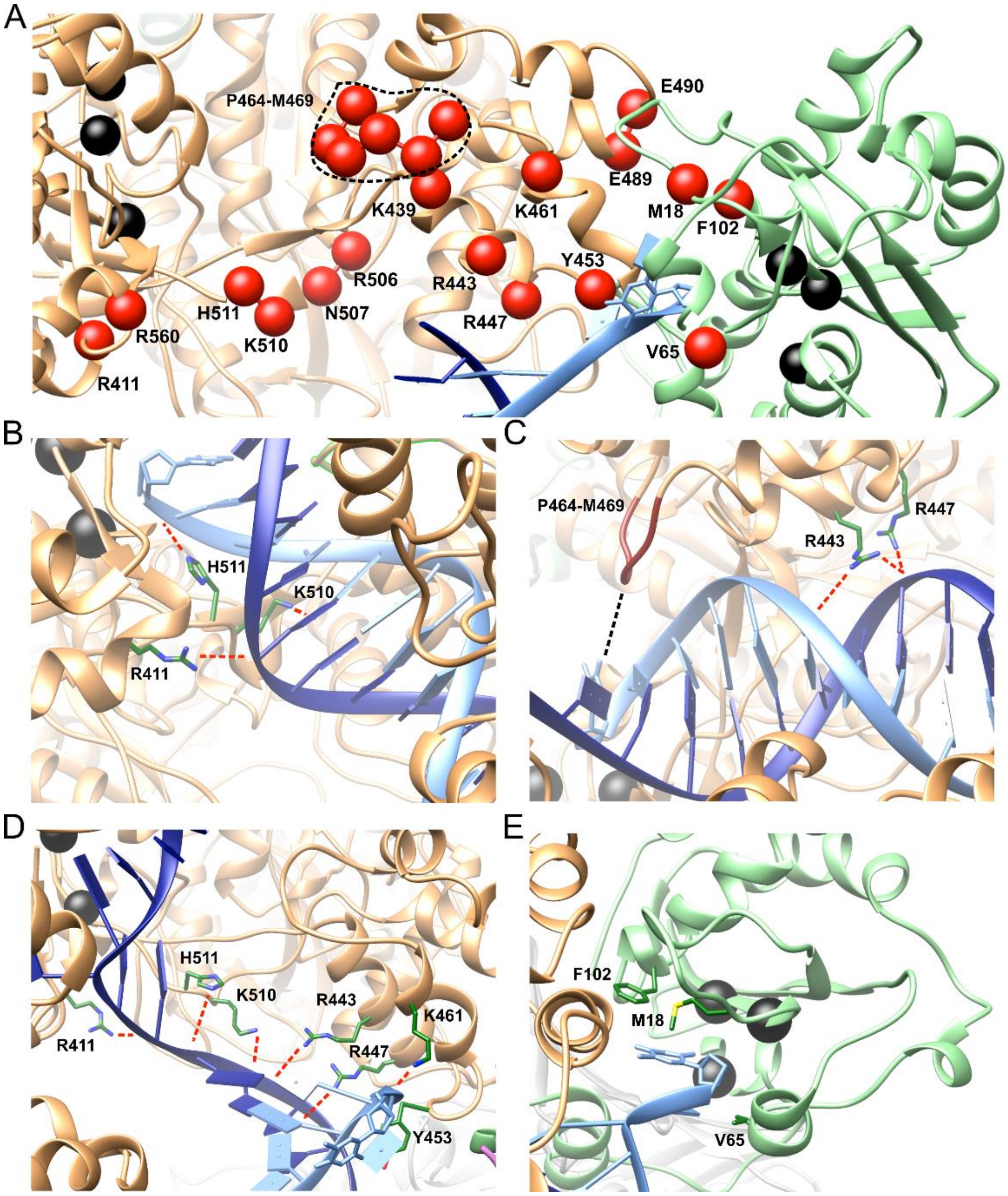

Figure 3.3.4 - Specific interactions along the optimal path accommodate the transitioning primer end to ensure facile pol-to-exo switching. A, Key residues (critical nodes) for pol-to-exo mode switching determined from dynamic network, conservation and persistent contacts analyses and mapped onto the Pol III structure. Critical nodes are shown as spheres, labeled and colored in red. Polymerase and exonuclease active site residues are shown as spheres and colored in black. B-D, Palm and thumb domain residues of the α subunit forming contacts important for polymerization (B, C) and for transitioning the primer end (D). Residue sidechains are shown in stick representation and labeled and colored by atom type (C is green, N is blue, S is yellow). Salt-bridge and polar interactions to the DNA are shown as dashed red lines. Hydrophobic interactions are shown as a dashed black line. E, Stabilization of the incoming mismatched nucleotide by the hydrophobic cluster of the ε subunit. Residues from the ε subunit hydrophobic patch are shown as sticks, labeled and colored in green.

3.3.4 Biochemical analysis of Pol III core mutants confirms an optimal transition path

To validate the defined path between the polymerase and exonuclease active site, we created eleven different mutations located in the vicinity of the polymerase and exonuclease active sites and along the path between the two sites. Seven of the mutations are located near the DNA in the palm or thumb domain (α439, α443, α447, α453, α461, α506/507, α510/511) (Fig. 3.3.5A). One of the mutations is located distal from the polymerase active site at the interface of the thumb and exonuclease (α489/490). Two are located in the exonuclease at the entrance of the active site (ε18, ε65) (Fig. 3.3.5B) A third exonuclease mutant ε102 was not soluble and therefore excluded from the experiments. Finally, we also deleted a loop in the thumb domain (residues 464-469: α loop) that protrudes into the DNA major groove, seemingly pushing it down into the polymerase active site, yet having no direct contact with the DNA[87].

All mutants were assembled into the trimeric polymerase-exonuclease-clamp complex and purified by gel filtration. To ensure a stable complex, an improved clamp-binder variant of the polymerase was used. This variant shows a >100-fold more stable complex than wild-type while retaining normal polymerase activity[87].

Next, we analyzed the polymerase and exonuclease activity of all the mutant polymerase-exonuclease-clamp complexes on different DNA substrates and conditions (Fig. 3.3.5). On a matched DNA substrate (containing a C:G base pair at the terminal position) (Fig. 3.3.5D) the two ε mutants and two α mutants located further away from the pol active site show no change in their activity compared with the wild type (ε18 and ε65, α loop, α489/490). The remaining α mutants show varying degrees of reduction in polymerase activity (α439, α443, α447, α461, α506/507) whereas two mutants are completely inhibited (α453, α510/511). None of the α mutations are part of the catalytic triad (composed of the three aspartates 401, 403, and 555)[99] but instead contact the DNA substrate backbone. Their reduced activity highlights the complexity of the polymerase active site and the necessity for accurate positioning of the DNA substrate for optimal polymerase activity.

Next, we tested the polymerase activity on a mismatched DNA substrate (containing a C:T mismatch at the terminal base pair), which requires the removal of the mismatched base before polymerase activity can proceed (Fig. 3.3.5E). The wild type and α mutants show no discernable difference in polymerase activity between the matched and mismatched substrate. In contrast, the exonuclease mutants ε18 and ε65 show an almost complete inhibition of activity on the mismatched DNA substrate. The ε mutants do not show significant difference in exonuclease activity when tested in isolation on a ssDNA, indicating that the reduction of activity is unique to the pol-exo-clamp complex, and the required transition of the primer strand from pol to exo site.

As the reduced polymerase activity of the majority of α mutants is masking the exonuclease activity in the DNA extension assay, we isolated the exonuclease activity from the polymerase activity by omitting dNTPs from the reaction conditions and followed DNA excision (Fig. 3.3.5F). On a matched (C:G) substrate, no exonuclease activity was observed, indicating that the DNA is prevented from reaching the exonuclease site. In contrast, on a mismatched substrate (C:T) the wildtype and several of the α mutants (α453, α510/511, α489/490, α loop) show robust removal of the first nucleotide. In all, only the first nucleotide is removed while the isolated exonuclease on ssDNA show processive exonuclease activity, further supporting the observation that within the pol-exo-clamp complex the polymerase protects matched DNA from the exonuclease. The remaining α mutants that are located in the path between the pol and exo active site (α439, α443, α447, α506/507) show decreased exonuclease activity on a mismatched DNA substrate, while α461 and the two ε mutants show a complete inhibition of exonuclease activity. As all the α mutant complexes are assembled with wildtype ε, and the ε mutants show near wild type activity on ssDNA, the reduced exonuclease activity of the pol-exo-clamp complexes indicates a defect in the transitioning of the primer strand from the polymerase to exonuclease active site. This is consistent with our MD analysis that predicted an essential role of these residues in the transfer of the primer strand between the two active sites.

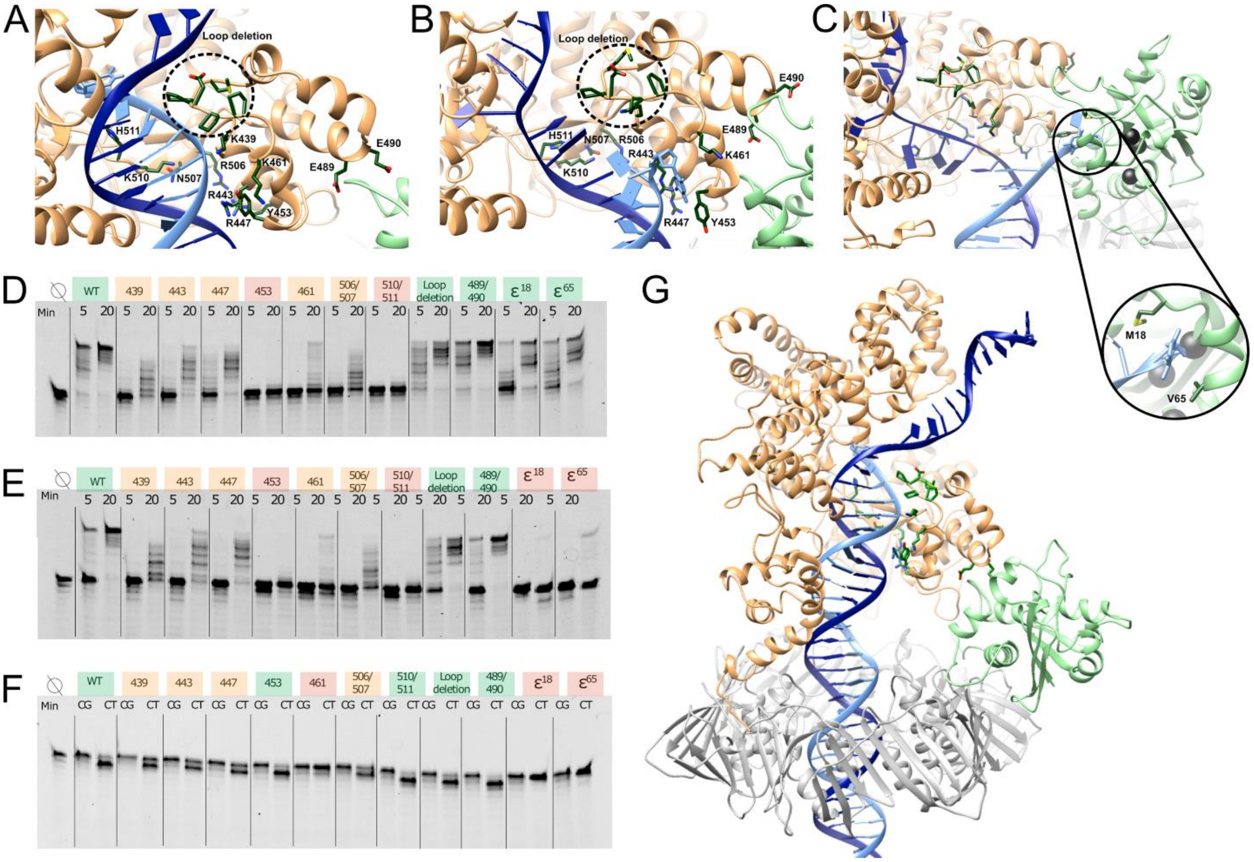

Figure 3.3.5 - Residues experimentally determined to be critical for transfer of DNA primer strand from polymerase to exonuclease active site. A-C, Close up of the transition path between polymerase and exonuclease active site in (A) polymerase mode (B) intermediate mode and (C) exonuclease mode. Polymerase is colored in orange, exonuclease in green, template DNA strand in dark blue, primer strand in light blue. Mutated residues are shown in dark green sticks. D, Denaturing gel analysis of polymerase activity of wild type and mutant proteins on matched DNA. Mutants showing W-T activity are highlighted in green, mutations that are moderately affected in orange, and mutations that render the protein inactive in red. E, Similar analysis using a DNA substrate with a terminal C-T mismatch. F, Exonuclease activity on matched 604 (C-G) and mismatched (C-T) DNA measured in the same DNA substrates as in panels D and E in the absence of nucleotides. G, Overview of Pol III core complex in polymerase mode. The mutated residues are highlighted in dark green and the β-clamp in gray. All experiments were performed by Margherita Botto.

3.4 Discussion

Replicative DNA polymerases achieve their remarkable fidelity by striking a delicate balance between DNA synthesis and excision of mis-incorporated nucleotides from the growing primer strand. To efficiently switch between DNA synthesis and excision, these versatile enzymes confine each activity into distinct active sites[119]. To ensure facile transfer of the DNA primer between these spatially separated sites, the entire DNA polymerase holoenzyme reorganizes along a well-defined conformational path. In this contribution, we combine state-of-the-art computational methods with closely coupled biochemical analyses to determine the optimal free energy path connecting the polymerization and exonuclease states of bacterial Pol III holoenzyme. We also use new data mining and classification strategies to discover kinetic intermediates, compute transition timescales and define molecular mechanisms based on analysis of the simulated Pol III conformational ensemble. Importantly, our results delineate a complete pol-to-exo mode switching mechanism addressing structural intermediates, protein dynamics, free energies and kinetics of the Pol III holoenzyme. All aspects of the mechanism emerge from our data analysis without a priori assumptions.

Our predictive mechanism involves stepwise melting of the first three nucleotides from the DNA primer end. Fraying of the mismatched terminal base pair is facile and occurs on a microsecond timescale at the earliest stages of the transition. The departure of the terminal thymine base from the polymerase active site results in a stalled polymerase state. Next, the Pol III holoenzyme exploits the natural motion of the DNA inside the Pol III/β-clamp central cavity to backslide and rotate, completely releasing the primer-template junction from the polymerase active site. The motion is guided by residues from the Pol III thumb and fingers domains. Base unpairing at the second and third position from the primer end is progressively more energetically costly.

Thus, the second and third unpairing events result in the highest barriers along the path and bring the overall timescale for the pol-to-exo transition into the millisecond range. High-energy intermediates on the landscape (plateau regions corresponding to S4 and S5) are stabilized by interactions with thumb domain residues, preventing backsliding toward pol mode. Notably, once base unpairing is complete the ssDNA primer undergoes fast sliding along the surface of the thumb domain, facilitated by contacts with strategically positioned residues. Consistent with their proposed mechanistic roles, mutations of these residues (e.g. α453, α461) slow down but do not abolish exonuclease activity. The fast rate of primer translocation compared to DNA melting has been noted in previous experimental studies and appears to be conserved across the A, B and C-family polymerases[114, 120-123]. The DNA primer's initial binding to the ε subunit is also a crucial step in ensuring efficient transfer. We identify a hydrophobic residue cluster that serves to stabilize the primer end prior to exonuclease site insertion. Importantly, we show that site mutations disrupting the cluster (ε18, ε65) affect exonuclease activity. Finally, we note the conformational shift and increased mobility of the ε subunit are essential features of our proposed mechanism. Insertion of the primer terminus into the exonuclease active site is gated by the motion of the ε subunit highlighting the important role of protein dynamics in the pol-to-exo mode transition.

In summary, our results shed light on the sophisticated strategies that allow replicative polymerases to achieve their extraordinary precision. Combining advanced computational modeling with insightful validation experiments, our study contributes to integrated understanding of high-fidelity DNA polymerases as dynamic assemblies engaged in safeguarding genome integrity.

3.5 Materials and Methods

3.5.1 Model Building and Equilibration

Pol III holoenzyme models (Pol III/β-clamp/primer-template DNA) we constructed in polymerization and editing modes from the available cryo-EM structures[86, 87]. The last base of the primer strand was converted to a thymine to produce a G:T' mismatch (' indicates primer strand). Missing residues in the ε flexible linker connecting the C-terminus with the catalytic domain were built using ModLoop[125]. The template strand was extended by 8-nt as the presence of an overhang has been shown to be important for exonuclease activity[86]. All simulations were carried out with the Amber ff14SB force field parameters[126] using the NAMD molecular dynamics code[127]. Electrostatics were calculated using the smooth particle mesh Ewald method and non-bonded interactions were evaluated with a 10-Å cutoff and 8.5-Å switching distance. Models were solvated with TIP3P water molecules from an equilibrated solvent box, ensuring 10 Å padding from the protein or nucleic acid atoms to the edge of the simulation box. Na+ and Cl-counterions were added to neutralize the overall charge of the Pol III holoenzyme complex and bring the ionic concentration to 150 mM. Both simulation systems were subjected to energy minimization for 5000 steps using the conjugate-gradient and line search algorithm and equilibrated for 5 ns with molecular dynamics flexible fitting (MDFF)[128] to ensure conformance to the respective cryo-EM densities. In the first stage of MDFF models were gradually heated to 300 K in the NVT ensemble while enforcing positional restraints on all heavy atoms using a force constant of 5.0 kcal mol^{-1} Å^{-2}. Positional restraints were then incrementally decreased to 0.0 kcal mol-1 Å^{-2} in the NPT ensemble (1 atm and 300 K). A scaling factor of $\xi = 0.1$ was employed during all stages of MDFF. By simultaneously decreasing positional restraints and enforcing weak

MDFF grid forces, both systems were allowed to gradually relax into their respective EM densities. The MDFF simulations employed a 1-fs timestep.

3.5.2 Path Optimization Protocol

A short 10-ns targeted molecular dynamics (TMD) run was used to connect the equilibrated end states. The exo-mode conformation was selected as the TMD target and the pol-mode conformation was driven to the target with a force constant of 1000 kcal mol^{-1} Å^{-2}. From the TMD trajectory we selected 32 evenly spaced snapshots (replicas) that served to initiate our path optimization protocol, employing the partial nudged elastic band method (PNEB)[51, 104]. The optimization protocol was carried out in several stages. First, replicas were heated to 300 K for 1.5 ns while employing a 20 kcal mol-1 Å-2 PNEB force constant. This was followed by a 3-ns run at 300 K using a 10 kcal mol^{-1} Å^{-2} force constant. For the subsequent 1.5 ns the chain-of-replicas was cooled to 0 K using a force constant of 20 kcal mol^{-1} Å^{-2}. This annealing cycle was repeated two more times to allow the replicas to gradually spread along the path and relax into local minima. The initial and final replicas were excluded from optimization to ensure conformance to the observed pol and exo conformations from cryo-EM. All protein and DNA heavy atoms were included in the PNEB calculation. The CUDA PMEMD module of the Amber molecular dynamics package was used for these simulations[129-131].

3.5.3 Unbiased and Biased Sampling Along the Minimum Energy Path

To sample extensively the conformational states along the optimal path, we initiated unbiased molecular dynamics simulations from each of 32 optimized replicas. Replicas were heated to 300 K for 500 ps in the NVT ensemble while imposing 5 kcal mol^{-1} Å^{-2} positional restraints on all heavy atoms. The restraints were scaled down to 0 kcal mol^{-1} Å^{-2} in a 5-ns NPT run. Each replica was then simulated for 200-ns using free unbiased MD, resulting in 6.4 μs of aggregate sampling along the PNEB path. All production runs were executed in the NPT ensemble (1 atm and 300 K) with a 2-fs timestep using the CUDA PMEMD module of the Amber code.

To improve sampling of regions in conformational space inaccessible to free unbiased MD we used umbrella sampling (US). These regions corresponded to barrier or high-energy plateau regions of the free energy landscape. We used a distance-based reaction coordinate (RC) for US biasing. Specifically, we selected the center-of-mass distance between the three nucleotides from the primer end and the exonuclease active site residues (D12, E14 and D167). The RC was subdivided into 12 overlapping windows with 1.0-Å spacing. Each window was simulated for 25 ns employing a force constant of 15 kcal mol^{-1} Å^{-2}. Umbrella sampling trajectories were then projected onto the first two eigenvectors obtained from time-lagged independent component analysis (TICA; refer to next section for details)[61, 62]. All umbrella sampling simulations were performed in the NPT ensemble (1 atm and 300 K) using the NFE module of AMBER. Configurations from the center of each umbrella window were then used as seeds for short (50-ns) unbiased MD simulations. This was done to ensure that the barrier regions contained both biased and unbiased sampling, a requirement for TRAM.

3.5.4 Time-Lagged Independent Component Analysis

To identify slowly varying degrees of freedom associated with the pol-to-exo conformational transition, we carried out dimensionality reduction on the trajectory data using time-lagged independent component analysis (TICA)[61, 62]. Atomic distances between the first ten base pairs of dsDNA and protein residues on the α/ε subunit were selected as collective coordinates for TICA. Residues from α/ε were selected by computing all protein contacts within 5 Å of the first ten base pairs of dsDNA across all configurations in the MEP. Phosphorous atoms on the backbone of each nucleotide and Cα atoms from each α/ε amino acid were used as a reference to compute the Euclidean distance between residues. Additionally, we included distances between N1 and N3 atoms on the first three base pairs that split and form the separated primer strand. In total, 456 unique distances were selected for dimensionality reduction with TICA. A lag time of $\tau = 500$ ps was used to compute the time-lagged covariance matrix. This matrix was then diagonalized to produce the respective eigenvectors and eigenvalues. MD trajectories were then projected onto the first two eigenvectors to yield the time-lagged independent components (ICs).

3.5.5 Multi-Ensemble Markov Model Estimation

Combining unbiased and biased simulation data allowed us to sample and achieve uninterrupted coverage of the transition path space defined by the first two ICs. K-means clustering was then employed in projected IC space producing 1000 microstate clusters. We then employed the transition-based reweighting analysis method (TRAM) to analyze our biased and unbiased simulation data producing correct free energy weighting of our microstates. A lag time of 500 ps was selected for the TRAM estimator based on the relaxation time of the estimated implied timescales. Kinetically similar microstates were then agglomerated into the S1-S8 macrostate clusters using the PCCA+ algorithm57. Mean first passages of times were then computed between macrostates using transition path theory[132] resulting in a kinetic model for primer translocation through the Pol III holoenzyme complex.

3.5.6 Bootstrapping

In order to compute error bars associated with the transition timescales and microstate free energies a bootstrap was performed. Prior to bootstrapping, multiple independent simulations were initiated from all 32 configurations along the MEP and from the center of each umbrella window. For each bootstrap sample one TRAM estimation was performed. Samples were generated by combining a simple bootstrap with a stationary bootstrap as described by Politis et al[133]. Under this paradigm, whole trajectories from the unbiased simulation data are drawn with replacement, while trajectory blocks of random length from the biased simulation data are drawn according to the stationary bootstrap algorithm. The minimum block length was selected to be the mean statistical inefficiency of the discretized trajectories in the umbrella sampling data set (5 ns). Error bars for the transition timescales and the microstate free energies in the barrier regions are presented in Table 3.5.1 and Table 3.5.2.

3.5.7 Critical Residues in the Pol III Holoenzyme Network

Dynamic network analysis was used to map protein and nucleic acid residues onto graphs wherein each residue is a node and contacting nodes are connected by edges (See Supplementary Fig. 6). All edges are weighted by dynamic correlation. Using these graphs, we computed suboptimal paths[116, 117] connecting the polymerization and exonuclease active sites for states S1-S8. Sampling of 50,000 frames from each macrostate were selected for this structural analysis. Suboptimal paths are a set of paths with length shorter than a specified limit above the optimal path. Suboptimal paths reflect residue correlations in molecular dynamics and, thereby, offer a way to quantify allosteric communication. Nodes traversed by the largest number of suboptimal paths frequently correspond to functionally important residues in the biological complex. Paths connecting αD403 and εD17 were calculated for macrostates S1-S8 using the Floyd-Warshall algorithm and a distance cutoff of 30 Å. We then normalized the distribution of critical residues across all macrostate suboptimal pathways. In addition to occupying privileged positions in the dynamic network we also require candidate residues to be conserved and to be in persistent contacts with the first ten dsDNA pairs for at least part of the pol-to-exo conformational transition. It was recently suggested that the E. coli polymerase is a phylogenetic outlier due to its ε-subunit[92]. Moreover, the E. coli-like exonuclease appears to exist explicitly in alpha-, beta- and gamma-proteobacteria leading us to only include these three groups in our conservation analysis. Amino acid conservation scores were determined using the EVfold server[134] and mapped to the structure of Pol III. Finally, we determined contacts between Pol III and the first ten dsDNA base pairs for all macrostates S1-S8. Protein residues were considered in contact if they were within 5 Å of the first 10 base pairs of the dsDNA. Contact persistence was computed as the frequency of appearance of the contact in the MD trajectories within the distance cutoff. Persistence values for each contact were then summed across all eight macrostates. Scores were obtained for each residue by combining their suboptimal path score, conservation score and contact persistence score. From the combined scores (Table 3.5.3) we selected the top 16 top scoring residues as candidates for experimental testing and validation.

3.5.8 Protein Purification and Complexes Assembly

All chemicals were purchased from Sigma Aldrich or Fisher Scientific, DNA oligonucleotides from Sigma and chromatography columns from GE Healthcare. Site direct mutagenesis was used to create nine mutants of the DNA Polymerase III α subunit and two mutants of the exonuclease ε (Supplementary Table 4). All proteins were expressed in E. coli BL21 (DE3). The α mutants were purified using a Histrap, Hitrap Q and a HiLoad Superdex 200 (120 ml) column. The β clamp was purified with a Histrap and a Hitrap Q column. The two exonuclease mutants ε18 and ε65 where purified from inclusion bodies in 6 M Urea using a Histrap column. The protein was then refolded by overnight dialysis into 0 M Urea and subsequently loaded on a Hitrap Q column. A third exonuclease mutant ε102 did not refold into soluble protein as was excluded from the studies. To assemble the complexes α, β and ε were mixed in a ratio 1:1.5:1.5, respectively, and loaded on a Superdex 200 Increase (2.4 ml) column. After 12% SDS PAGE gel analysis, fractions that contained the three proteins were pooled together. The individually created mutant complexes were further analyzed by SDS Page using 4-20% Mini-PROTEAN TGX Precast Protein Gels to confirm all complexes were at the same concentration. All proteins and complexes were flash frozen in liquid nitrogen and stored at -80 °C.

3.5.9 DNA Primer Extension Assay

Polymerase and exonuclease activities were measured using a 26 base pair dsDNA substrate with a 11-nucleotide single stranded overhang. (template strand: 5' GCTAGCTTACACGAGTCCTTCGTCCTAGTACTACTCC; matched primer strand: 5' 6-FAM GGAGTAGTACTAGGACGAAGGACTCG 3'; mismatched primer strand: 5' 6-FAM GGAGTAGTACTAGGACGAAGGACTCT 3'). All the reactions were performed at room temperature in 20 mM Hepes pH 7.5, 2 mM DTT, 5 mM MgCl2, 50 mM NaCl and 0.5 mg/ml BSA. For the experiments with nucleotides 100 μM dNTPs (each) were added to the buffer. Reactions were started by addition of 40 nM of protein complex to 50 nM of DNA (final concentrations). Reactions with dNTPs were stopped at 5 and 20 minutes, while reactions without dNTPs were stopped at 5 minutes. All the reactions were performed with matched (CG) and mismatched DNA (CT). Reactions were then run on a denaturing 20% Acrylamide (19:1) gel in 1xTBE with 6M Urea for 1 hour and 20 minutes at 30W. Afterwards the gel was imaged on Typhoon using Alexa Fluor 488 filter.

Table 3.5.1 – Associated errors of the transition timescales determined with TRAM.

Macrostate Transition	Timescale (µs)	StdErr +/- (µs)
S1 → S2	1.9	0.08
S1 ← S2	0.4	0.01
S2 → S3	19.8	0.03
S2 ← S3	3.9	0.02
S3 → S4	17.3	0.01
S3 ← S4	0.03	0.001
S4 → S5	2100	1.2
S4 ← S5	90.1	0.1
S5 → S6	5100	2.2
S5 ← S6	13300	12.8
S6 → S7	0.2	0.03
S6 ← S7	0.3	0.01
S7 → S8	711.1	0.8
S7 ← S8	1400	1.3

Table 3.5.2 – Associated errors of the microstate free energy estimates determined with TRAM

Macrostate	ΔG (kcal/mol)	StdErr +/-
S4:S5	9.33	0.32
S4:S5	10.41	0.37
S4:S5	10.96	0.43
S5:S6	13.14	0.79
S5:S6	15.92	0.66
S5:S6	15.83	0.26
S5:S6	15.54	0.29
S5:S6	16.01	0.56
S5:S6	16.18	1.02

Table 3.5.3 – Critical residue scores determined from combining suboptimal paths, conservation and contact persistence data.

Residue	Subopt Distribution	Conservation Score	Summed Contact Persistence	Total
R411	0	0.84	1.2	2.0
K439	0.20	0.50	3.8	4.5
R443	0.35	0.67	2.4	3.4
R447	0.29	0.48	1.6	3.0
Y453	0.35	0.45	2.8	3.7
K461	0.54	0.35	1.8	2.7
E489	0.90	0.38	0.0	1.3
E490	0.68	0.45	0.0	1.1
R506	0.10	0.94	1.3	2.3
N507	0.17	0.42	1.1	1.7
K510	0.15	0.32	5.8	6.3
H511	0.05	0.56	4.1	4.7
R560	0	0.80	4.0	4.8
M18ε	0.41	N/A	1.8	2.2
V65ε	0.37	N/A	1.7	2.1
F102ε	0.25	N/A	1.7	2.0

CHAPTER 4. UNCOVERING UNIVERSAL RULES GOVERNING THE SELECTIVITY OF THE ARCHETYPAL DNA GLYCOSYLASE TDG

4.1 Abstract

Thymine DNA glycosylase (TDG) is a pivotal enzyme with dual roles in both genome maintenance and epigenetic regulation. TDG is involved in cytosine demethylation at CpG sites in DNA. Here we have used molecular modeling to delineate the lesion search and DNA base interrogation mechanisms of TDG. First, we examined the capacity of TDG to interrogate not only DNA substrates with 5-carboxyl cytosine modifications but also G:T mismatches and non-mismatched (A:T) base pairs using classical and accelerated molecular dynamics. To determine the kinetics, we constructed Markov State Models (MSM). Base interrogation was found to be highly stochastic and proceeded through insertion of an arginine-containing loop into the DNA minor groove to transiently disrupt Watson-Crick pairing. Next, we employed novel path sampling methodologies to compute minimum free energy paths for TDG base extrusion. We identified the key intermediates imparting selectivity and determined effective free energy profiles for the lesion search and base extrusion into the TDG active site. Our results show that DNA sculpting, dynamic glycosylase interactions and stabilizing contacts collectively provide a powerful mechanism for the detection and discrimination of modified bases and epigenetic marks in DNA.

4.2 Significance Statement

The most prominent epigenetic modification in mammalian genomes is cytosine methylation at position 5 on the pyrimidine ring. Thymine DNA glycosylase plays a central role in the pathways for 5-methyl cytosine removal and, thus, influences gene silencing, stem cell differentiation and alterations in normal development. Additionally, methylation abnormalities in DNA are often observed in diseases, specifically cancer. Here we examine the mechanisms by which TDG detects, extrudes and excises modified bases in DNA. Using novel path sampling methodologies, we compute minimum free energy paths for TDG base extrusion. The computed paths reveal a novel mechanism underpinning TDG selectivity for DNA lesions or modified bases, which involves DNA sculpting, global protein dynamics, conformational gating and specific protein-nucleic acid interactions.

4.3 Introduction

Genome maintenance occurs in the context of chromatin and it is becoming increasingly apparent that epigenetic regulation is intricately intertwined with the DNA damage response in ensuring genome stability. Understanding how epigenetic marks are recognized, distinguished from exogenous or endogenous DNA lesions, and processed by the canonical DNA repair machinery is a topic of great current interest. Here our focus is on the base excision repair (BER) pathway, which in addition to an established role in genome maintenance, is associated with many other cellular processes[135], including a recently discovered critical role in epigenetic regulation[136-139]. The most prominent epigenetic modification in mammalian genomes is cytosine methylation, which typically occurs at CpG islands and enhances chromatin packing to promote gene silencing[140]. Consequently, 5mC demethylation is crucial for resuming the transcription of silenced genes. Notably, unbalanced cytosine methylation is a hallmark of cancer[141-143]. In cancer, predominantly demethylated regions of the genome could become hyper-methylated leading to the silencing of tumor suppressor genes. Furthermore, 5-methyl cytosine deamination results in G·T mismatches that could cause C to T transition mutations during DNA replication. It is estimated that nearly a third of cancer mutations found in coding regions of the genome arise from C and 5mC deamination at CpG sites[137]. There is also a clear link between aging and methylation levels in GpG islands[144]. The importance of maintaining the methylation state of the genome requires tight regulation of pathways controlling the levels of 5mC. Removal of 5-methylcytosine (5mC) bases (Figure 4.3.1) is known to proceed through successive steps of oxidation by enzymes from the ten-eleven-translocase (TET) family, producing 5-formyl cytosine (5fC) and 5-carboxyl cytosine (5caC) intermediates[140, 145, 146]. Unlike methyl cytosine, these intermediates are substrates for thymine DNA glycosylase[140, 147, 148] – a classic DNA repair enzyme. While TDG is important for the repair of mutagenic DNA lesions, it has an even more prominent role in ensuring epigenetic stability. In this capacity, TDG activity is vital during embryonic development[149]. TDG also interacts with numerous protein partners engaged in epigenetic regulation (e.g. DNMT3a[150] and CBP/p300 acetylase[151]) and transcription (transcription factors, nuclear receptors [152]) and is intricately involved in the regulation of gene expression.

A second pathway to process 5mC is through deamination followed by the action of MBD1-4 glycosylases[153, 154]. The resulting abasic DNA is then channeled through the BER pathway. BER efficiency relies on a remarkably discriminating search for modified bases among an enormous background of normal DNA. The search is followed by damage-specific base extrusion into to the enzyme’s active site, removal of the damaged bases and handoff of the product DNA to downstream pathway participants.

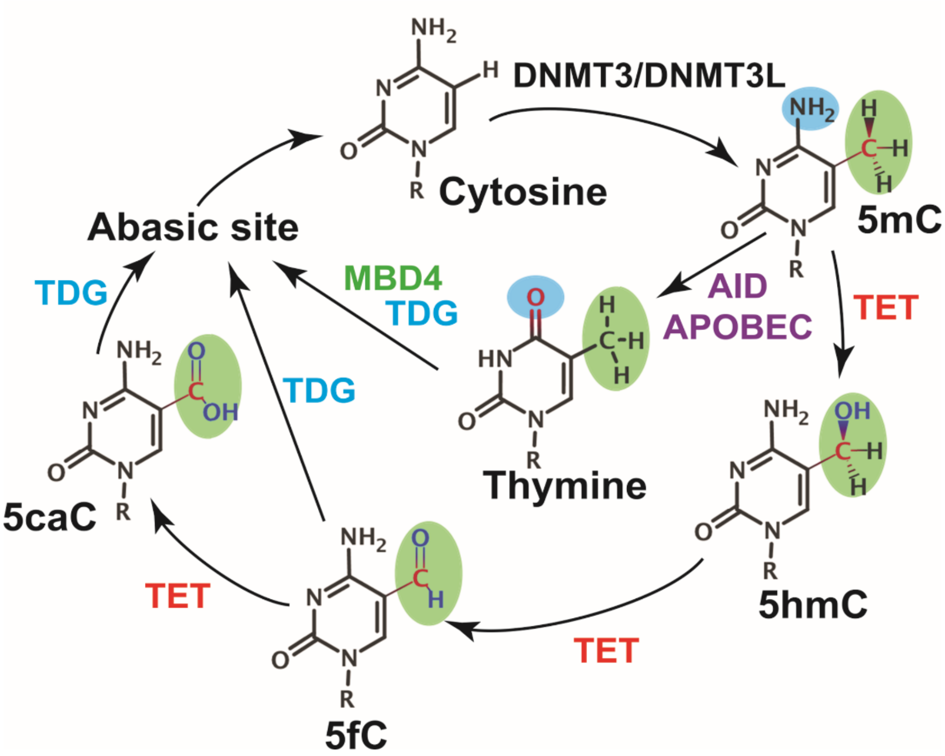

Figure 4.3.1 - Schematic representation of the two active demethylation pathways known in mammals. MBD4-mediated pathway is shown with blue arrows; the TET-mediated pathway is shown with red arrows.

Here we establish a basis to understand the key principles underpinning the extraordinary power of the TDG glycosylase to discriminate in favor of modified bases against a backdrop of normal genomic DNA. We further elucidate the protein-nucleic acid interactions ensuring specificity for lesions or epigenetic marks. Key to selectivity is nucleotide extrusion - a process involving a nucleotide swinging out of the DNA helix and being accommodated in the catalytic pocket of TDG. Nucleotide extrusion[155-157] is a major determinant of glycosylase selectivity, with potential for selection or rejection of substrates at each intermediate along the base eversion path. Glycosylases[156, 158] also employ DNA sculpting strategies (e.g. DNA bending and loop insertion) to lower the energetic barrier of base extrusion and, thus, increase the efficiency of the dynamic lesion search. Whether glycosylases employ active or passive strategies in this search process has been a topic of considerable debate. NMR evidence has suggested glycosylases could act as passive kinetic traps for spontaneously exposed extrahelical bases[159, 160]. Conversely, evidence from molecular crystallography (MX) has pointed to active base extrusion mechanisms. Numerous glycosylase structures[154, 156, 161-163] have shown that DNA binding is accompanied by a multitude of conformational changes preceding active site chemistry: 1) DNA sculpting through interactions with the enzyme DNA-binding groove; 2) DNA bending, minor groove compression and backbone distortion at the lesion site; 3) residue insertion into the DNA stack to expel the lesion base and stabilize the orphaned base; and 4) base flipping into a lesion-specific recognition pockets that sterically exclude non-lesion bases. Crosslinking strategies have, in rare instances, captured crystallographic snapshots of base extrusion intermediates[164] and could, in principle, provide insight into short-lived species along base extrusion paths. Nonetheless, base flipping is inherently dynamic and, therefore, not easily construed from static crystallographic snapshots. Therefore, molecular modeling studies have been extensively used to complement structural biology approaches and have proven enormously valuable in unraveling detailed dynamics of glycosylase enzymes and the origins of selectivity in BER[157, 163, 165].

4.4 Results and Discussion

To explore whether TDG employs an active or passive mechanism, we carried out simulations on TDG/5caC-DNA complexes. As starting points for computational modeling, we utilized existing structures of TDG/5caC-DNA in a pre-extrusion and post-excision state (PDB codes: 2RBA and 5HF7)[166, 167]. The following initial models were created: (i) pre-extrusion state (5caC accommodated in the DNA base stack); (ii) fully extruded state (5caC inserted in the TDG active site); and (iii) an initial interrogation complex. To address base interrogation, we started from systems with initially separated TDG and 5caC-DNA and simulated complex formation. We also simulated TDG in the presence of a G:T mismatch and with normal DNA.

Our first goal was to delineate the accessible conformational space for TDG/5caC-DNA and to assess the capacity of TDG to interrogate not only 5caC but also G:T mismatches and non-mismatched (A:T) base pairs. To this end, we carried out accelerated MD[168] (aMD) runs on the TDG interrogation complexes. The aMD method enhances sampling of the torsional degrees of freedom to accelerate phase space exploration and facilitate transitions over high energy barriers. Surprisingly, we observed that the presence of TDG induces multiple transient base-opening events that occur within 200 ns of aMD sampling. Base interrogation was found to be highly stochastic and appeared to proceed through insertion of an arginine-containing loop (Arg275) into the DNA minor groove to transiently disrupt Watson-Crick pairing. To ascertain that these events are also detectable in unbiased MD, we carried out multiple trajectory regular MD runs for an aggregate simulation time of 8 μs. Analogous runs were performed on the G:T mismatch and unmodified DNA systems. DNA backbone torsion angles for the interrogated base pair and distances between the two bases and the guanidinium group of Arg275 were selected as coordinates for time-lagged independent component analysis (TICA)[62, 169]. The combined trajectories were projected onto the first two independent components (ICs) of the A:T system, which allowed for direct comparison of all three substrates. Different energy minima (metastable states) are present and clearly separated in the TICA projections (Figure 4.4.1A-C). All trajectory frames were then clustered in the projected space of the two ICs using the k-means algorithm, producing 800 clusters (i.e. microstates). From this data we constructed Markov State Models (MSM)[71, 110] and evaluated the kinetics of TDG base interrogation using transition path theory[170]. Results are presented in Figure 2D-F. Several conclusions are immediately apparent from our analysis. First, TDG probes DNA bases non-specifically, interrogating not only 5caC-modified bases but also on G:T mismatches and normal base pairs. Among the identified kinetically distinct macrostates we distinguish two low-populated extrahelical states with Arg275 inserted into the DNA stack in two different orientations (Figure 4.4.2). These states are accessed through two kinetic intermediates: (1) an intermediate with TDG-induced local torsional shift and intact Watson-Crick pairing; and (2) an intermediate with partially broken Watson-Crick pairing and Arg275 inserted between the extruded and the orphaned base (Figure 4.4.2C and Figure 4.4.2D). The second key observation is that the extrahelical states are extremely short-lived, and thus, not readily detectable by NMR (Figure 4.4.2A). The free energy landscapes in Figure 4.4.2 reflect this, with barriers to the extrahelical states not exceeding 4 kcal/mol. Third, the H-bonding between Arg275 and 5caC is variable and differs from the pattern for normal DNA (Figure 4.4.3). This observation rationalizes the differences between the free energy landscapes in Figure 4.4.1A and 4.4.1C with a lower barrier for a 5caC modified base to access the extrahelical state.

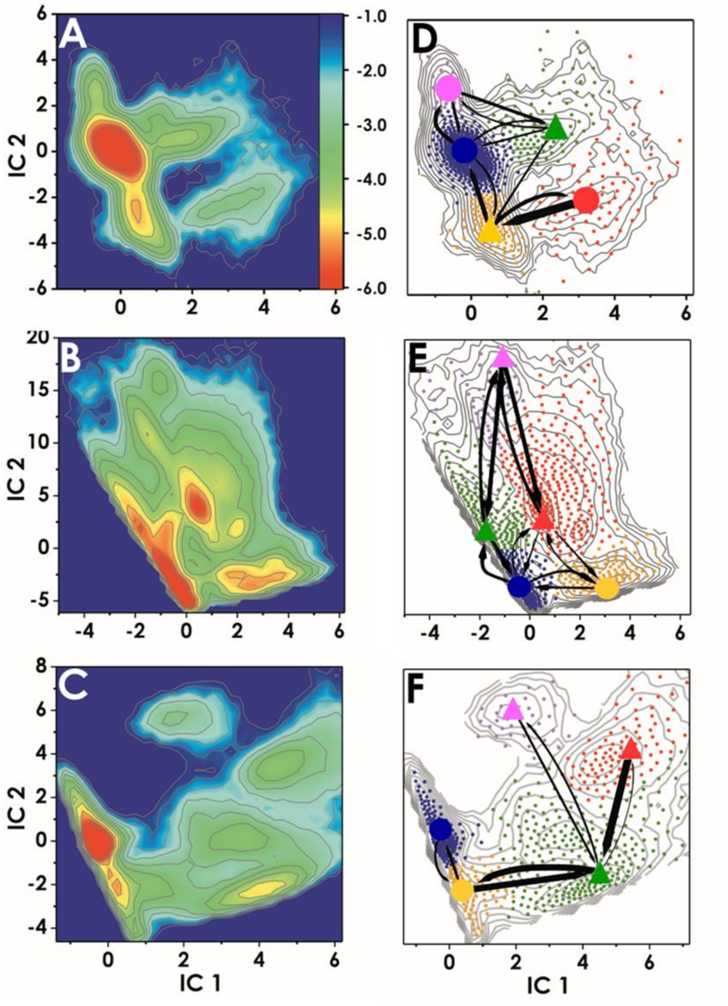

Figure 4.4.1 - Conformational dynamics of TDG/DNA complexes during base interrogation. Computed free energy profiles projected onto the first two ICs for (A) 5caC-DNA (B) G:T mismatch and (C) A:T base pair. The color bar inset denotes the ΔG scale in kcal/mol. Results from MSM analysis for (D) 5caC-DNA (E) G:T mismatch and (F) A:T base pair systems. Panels D-F show the positions of all microstate clusters in the space of the two ICs (small dots); how these small clusters were agglomerated into macrostates (denoted by large dots for the intrahelical or triangles for the extrahelical states) and the probabilities of transitions between macrostates. Microstates (dots) are colored by the macrostate they belong to and the coloring scheme is consistent between all three systems. The relative thickness of the arrows connecting the macrostates denotes the macrostate transition probabilities computed using transition path theory.

We also analyzed DNA structural parameters using the Curves+ code[171] to determine if TDG exploits local DNA deformation to facilitate lesion interrogation and selection. We found very little difference between the inter-base parameters as well as the total bend of the 5caC and the A:T intrahelical states (Table S5). However, analysis of the transient extrahelical states revealed changes to both the shift and tilt values (~2 Å and ~6.5◦) at the interrogation site and flanking base of 5caC. Importantly, the negative charge on 5caC provides a convenient handle for TDG to stabilize an extrahelical intermediate using an opportunely positioned Lys201 (Figure 4.4.3E).

G:T mismatches could also rapidly transition between non-extruded and extruded states (Figure 4.4.1E), exhibiting two well-defined metastable states with disrupted base pairing. This can be rationalized by the fact that G:T mismatches form "wobble" hydrogen-bonding pairs, which require a sideways shift of one base relative to Watson-Crick positioning. Our structural analysis of the interrogation site confirms this, with intrahelical shift and twist values differing significantly (by ~2 Å and ~5◦, respectively) from the ones measures for the intrahelical states of the 5caC and A:T. This leads to increased propensity for bending and base pair disruption at the G:T site[172-175], which TDG takes advantage of through backbone distortion alone. Thus, TDG has shown stronger G:T mismatch repair activity, in vitro, when compared to modified substrates[176]. Interestingly, Curves+ analysis of the interrogation states for the G:T mismatch system indicates that base stacking is disrupted not only for the transient extrahelical macrostates but also for the intrahelical basin that contains the majority of conformers from the unbiased MD trajectories. Thus, unlike the regular (A:T) DNA or 5caC, the G:T mismatch has a local structural distortion at the very outset of interrogation. This leads to an energetically destabilized initial state and lower energetic cost for base extrusion in the G:T mismatch system.

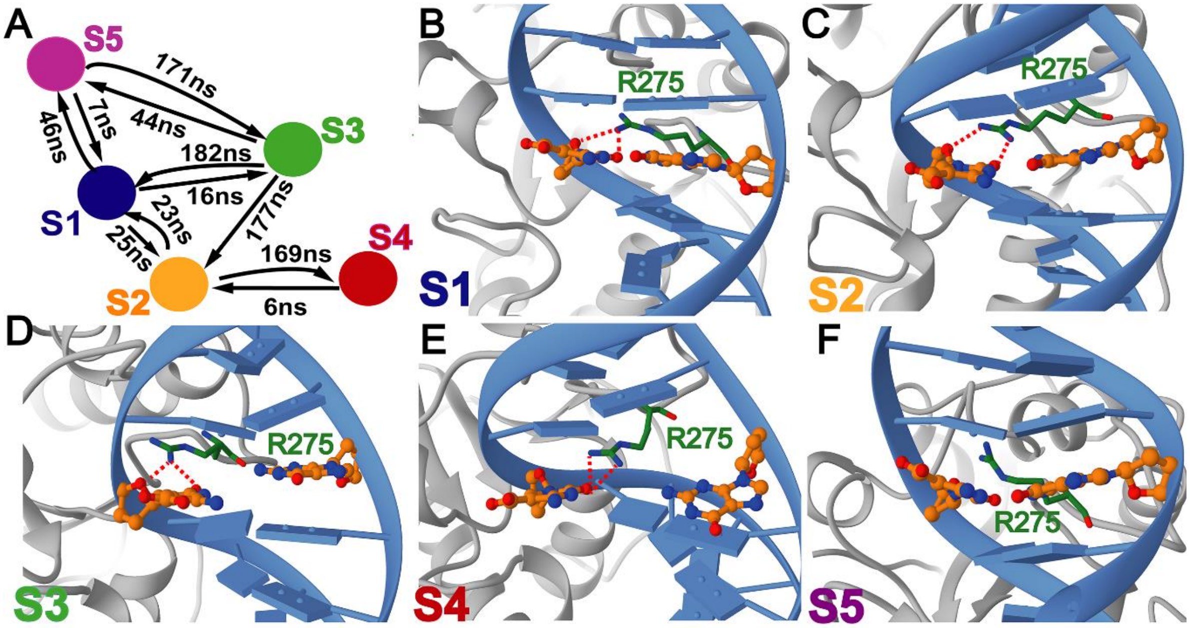

Figure 4.4.2 - Representative structures selected from each macrostate of the Markov State Model corresponding to 5caC base interrogation by TDG. (A) Calculated transition timescales determined from transition path theory. Representative structure for macrostates (B) Intrahelical state S1, (C) Extrahelical state S2, (D) Extrahelical state S3, (E) Partially extruded state S4 and (F) Intrahelical state S5. Each state is colored according to the color scheme in Figure 4.4.1D; panels are labeled by macrostate designation. TDG is shown in grey; DNA is shown in blue. The intercalating Arg275 residue at the tip of the insertion loop is shown in ball and stick representation and colored in green. The extruded 5caC base and the orphaned base are shown in ball and stick representation and colored in orange.

These results further substantiate the differences in the transition timescales, with G:T base extrusion occurring on a faster timescale than both the 5caC and A:T substrates (Figure 4.4.2).

The results also highlight the role of the TDG insertion loop (Ala270-Pro280) and particularly Arg275, which stabilizes the extrahelical intermediates by stepwise replacement of the Watson-Crick hydrogen bonds. Several glycosylases have been proposed to utilize similar mechanisms in which both DNA sculpting and loop insertion is exploited. Structures of hOGG1, its bacterial homolog MutM, and MBD4 all exhibit Arg-loop insertion[154, 163, 177, 178]. Similarly, single molecule experiments have shown the E. coli repair enzymes Fpg, Nei and Nth to utilize DNA sculpting and intercalating loop strategies to interrogate and extrude damage bases[179, 180].

As the next step in our analysis, we determined the complete base eversion path for TDG/5caC-DNA, starting with the interrogation complexes and ending with the fully extruded state. We identified the key intermediates imparting selectivity and also computed an effective free energy landscape for this transition. Our results show that base eversion in TDG is a gated process that involves motions of several flexible loops and a gating helix (Figure 4.4.3). Therefore, intuitive reaction coordinates (e.g. pseudo torsions) are, in this case, not practical. Recently, there has been considerable progress in methods[181-183] to optimize minimum energy paths (MEP) when the initial and final states are known. We leveraged two of these methods, the partial nudged elastic band (PNEB)[184, 51] and finite temperature string method[49, 50], to investigate recognition of modified bases by TDG. Both methods define the MEP as a chain of replicas of the system connecting the initial and final configurations. First, we optimized a MEP between the pre-extrusion and the fully extruded states using PNEB. Gradually spreading the replicas from these two states allowed the optimization process to discover the path in an unbiased way.

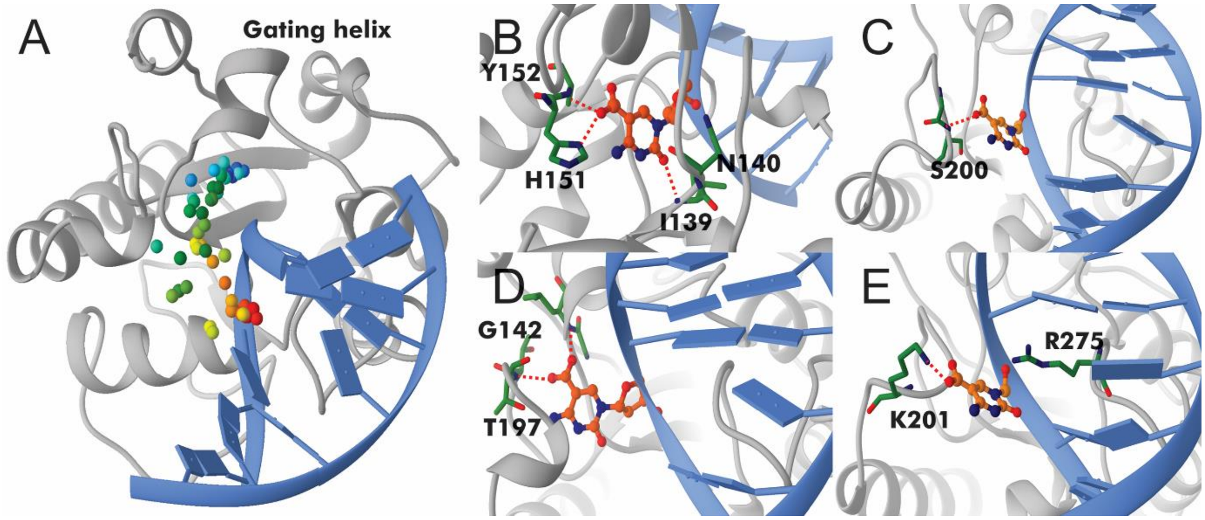

Figure 4.4.3 - Optimized base extrusion path for 5caC in TDG. (A) The path is indicated by colored dots tracking the position of the carbon atom of the 5caC carboxylate group along the path. Color denotes the replica index from initial (red) to final (blue). Four snapshots along the path are shown. (B) Early intermediate with Arg275 intercalating into the DNA; (C) 5caC at the Pro198 loop; (D) 5caC positioned next to the gating helix; and (E) 5caC extruded into the TDG active site.

The PNEB optimized path served as a starting point for further optimization with the finite temperature string method, which could provide more extensive path sampling. Since the string method works in projected collective variable (CV) space, a preliminary PNEB step was necessary to provide an unbiased initial path and to select CVs for the string method. Using this protocol for TDG/5caC-DNA, we completed 25 ns of PNEB optimization (with 28 replicas) and 224 ns (200 iterations) of the string method (Figure 4.4.3). After MEP convergence, we released each replica and sampled an aggregate of 11.2 μs of unrestrained MD trajectories along the base eversion path, which we further analyzed to construct an MSM.

Our results reveal an intricate network of protein-DNA contacts necessary to accommodate the 5caC base during its passage from the DNA base stack into the TDG active site. Importantly, these contacts significantly lower the free energy barriers for base extrusion to approximately 4 kcal/mol (Figure 4.4.4A). By comparison, umbrella sampling simulations of base eversion in the absence of the glycosylase result in barriers of at least 12 kcal/mol, which is consistent with previously published values for base extrusion barriers in DNA[155-157]. From the MSM analysis we identify 6 kinetically distinct states along the 5caC eversion path (Figure 4.4.4B, 4.4.4C and Figure 4.4.5). State S1 has the modified base accommodated in the stack; state S2 is an early intermediate wherein 5caC is inserted between the intercalating Arg275 and Lys201. The positive charge on the lysine stabilizes the negative charge on the 5caC carboxyl group. States S3, S4 and S5 correspond to configurations wherein 5caC interacts with residues of the Pro198 loop of TDG. Indeed, access to these three relatively rapidly interconverting states is gated by the motion of the Pro198 loop and the adjacent helix (Ser205-Lys221). The TDG gating helix (Leu143-Lys148) serves as a secondary gate by closing over the active site after the 5caC base is inserted. Thus, base eversion by TDG is a global conformational transition and conformational gating is necessary for the 5caC base to access the active site.

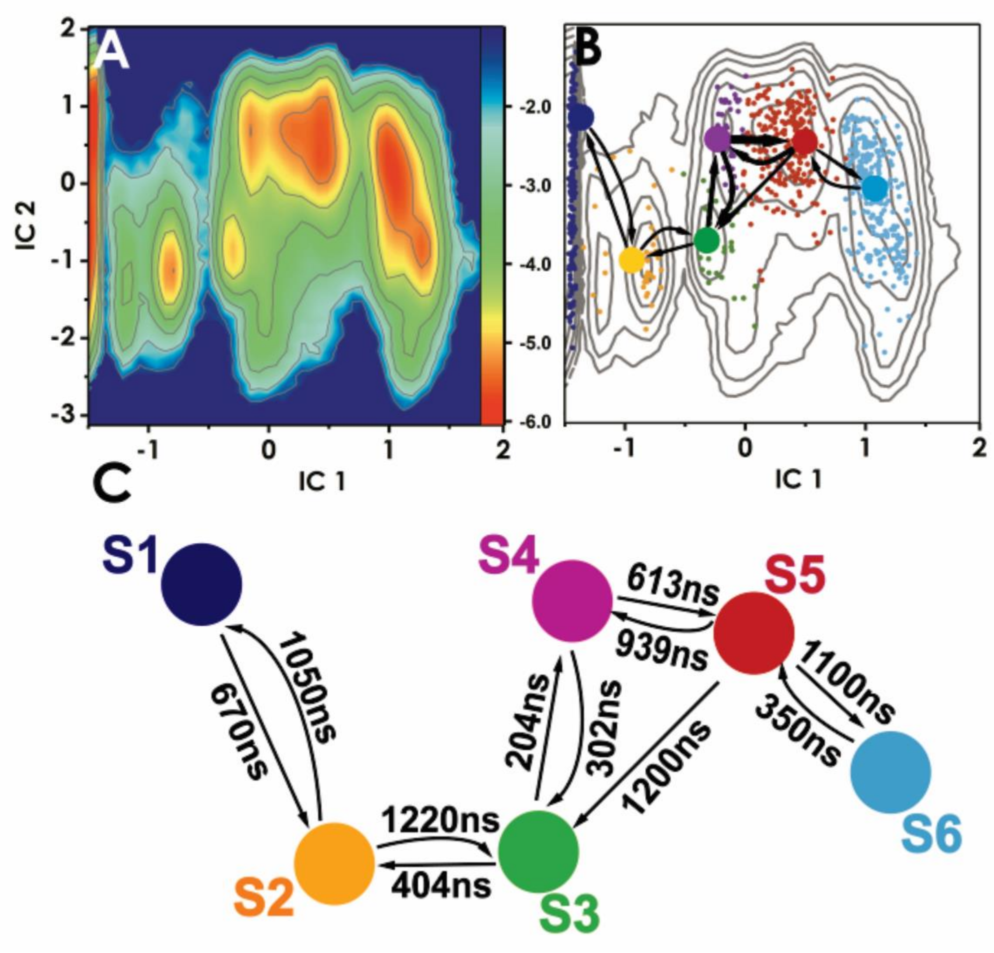

Figure 4.4.4 - Conformational Dynamics of TDG/5caC-DNA complex during base eversion. (A) Free energy profile projected onto the first two ICs. Color bar inset denotes ΔG scale in kcal/mol. (B) Results from MSM analysis. Microstates (dots) are colored by macrostate they belong to. Probability fluxes between macrostates from transition path theory are shown by arrows. (C) Calculated macrostate transition timescales.

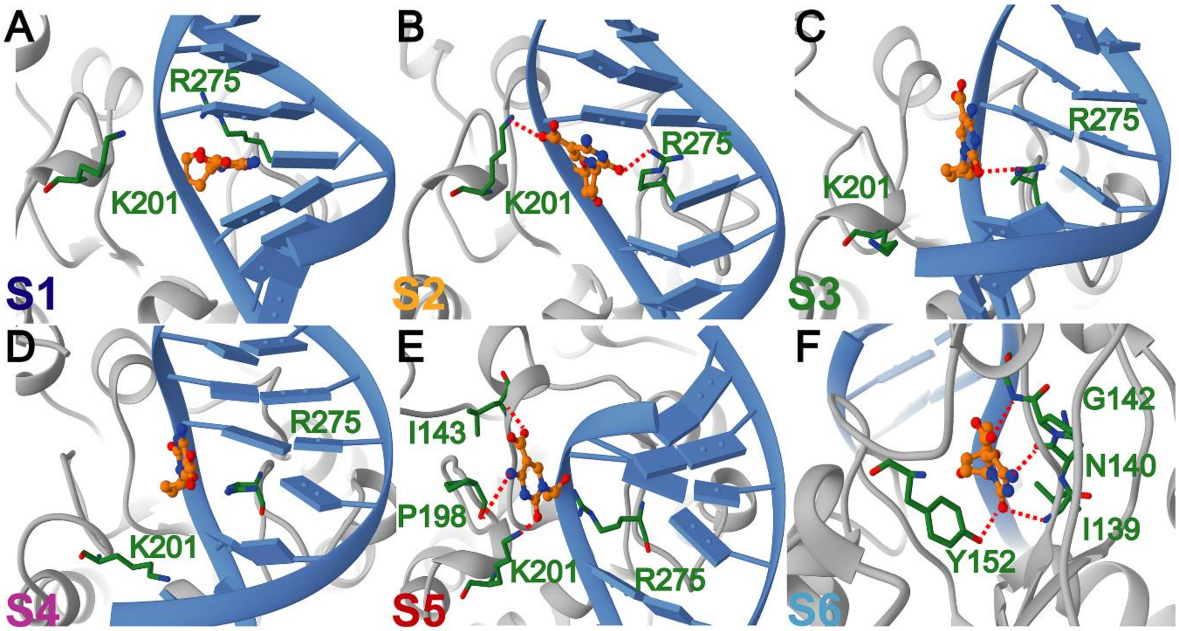

Figure 4.4.5 - Representative structures selected from each macrostate along the optimized TDG base eversion path for 5caC DNA. (A) Intrahelical state S1, (B) Partially extruded state S2, (C) Extrahelical state S3, (D) Extrahelical state S4, (E) Extrahelical state S5 and (F) Extrahelical base in TDG active site, denoted S6. Each state is colored according to the color scheme in Figure 4.4.4B; panels are labeled by macrostate designation. TDG is shown in grey; DNA is shown in blue. Residues contacting the extruded base are explicitly shown and labeled in green. The 5caC base is shown in ball and stick representation and colored in orange. Hydrogen bonds to the extruded base are denoted as red dash lines.

4.5 Conclusions

Our computational modeling reveals a novel mechanism underpinning TDG selectivity for DNA lesions (G:T mismatches) or modified bases (e.g. 5caC), which involves DNA sculpting, global protein dynamics, conformational gating and specific protein-nucleic acid interactions that stabilize the extruded base along the path from the DNA stack to the TDG active site. Our model for base extrusion by TDG bears certain similarities to an earlier proposed mechanism ("pinch-push-pull") for human UNG [185-188]. In this model, the "pinch" involved compression of the DNA backbone such that the distances between the phosphates flanking the uracil base were reduced by ~4 Å. Three static enzyme loops were proposed to mediate DNA recognition: the minor groove reading loop (His268-Ser273), the Pro-rich loop (Pro165-Pro168) and the Gly-Ser loop (Gly246-Ser247). Nucleotide flipping was proposed to be facilitated by the intercalation ("push") of Leu272 into the DNA base stack. The final step was the pulling of the uracil base and ribose ring deep into the uracil recognition pocket, resulting in hydrogen bonding to every polar atom of the uracil and in face to face π stacking with Phe158 and Tyr147. In a similar scenario for TDG, the "pinch" step is achieved by DNA sculpting via protein-DNA interactions and dynamic Arg275-loop insertion. This leads to kinking of the DNA substrate and compression of the distance between the flanking phosphates above and below the extrusion site by up to 3 Å in the extrusion intermediates (Table 5.5.1). The "push" involves insertion of an interrogation loop into the DNA minor groove, intercalation of an arginine (Arg275) from the tip of the interrogation loop into the DNA stack, and stepwise replacement of Watson-Crick H-bonds to lower the energetic barrier for base flipping. Finally, the "pull" step is achieved by accommodation of the extrahelical base via specific residue interactions in four stable intermediates along the extrusion path (Figure 4.4.5). However, there are also important differences with the previous model. First, unlike Leu272 in

UNG, which plays a role of a steric plug, Arg275 has the capacity to actively disrupt Watson-Crick hydrogen bonding. Stepwise replacement of Watson-Crick H-bonds between the extruded and orphaned base lowers the barrier to reach the most populated extrahelical state during interrogation. Second, the “pinch-push-pull” model originated from molecular crystallography and emphasized the role of static enzyme loops. By contrast, we show that protein dynamics and global gating motions of TDG are essential. Specifically, transitioning the 5caC base into the active site requires gating motions of the Pro198 loop and the adjacent helix (Ser205-Lys221) as well as motions of the TDG gating helix (Leu143-Lys148). These motions cannot be easily construed from static crystal structures. Collectively, our results shed light on the key determinants of glycosylase selectivity and uncover universal rules governing this class of enzymes.

4.6 Materials and Methods

4.6.1 Model Construction

Models for the pre- and post-extrusion states were constructed from two TDG/DNA crystal structures (PDB ID: 5HF7[166] and 2RBA[167]). For the base interrogation, we built the system with initially separated TDG and 5caC-DNA. We also built TDG-DNA complexes with a G:T mismatch and normal DNA (A:T pair). For consistency, all systems were built with the same DNA sequence 5’-GTACGTGAG-3’. All systems were then solvated with TIP3P[29] water molecules in a box with a minimum distance of 10.0 Å from the surface atoms of the complex to the edge of the periodic simulation box. Counter-ions were added to neutralize the net charge of the complex and reach 150 mM NaCl concentration to mimic physiological conditions. 5caC force field parameters were determined with the Antechamber module of AMBER16[131].

4.6.2 Molecular Dynamics for Base Interrogation

Steepest decent minimization was performed for 5000 steps. Each system was then slowly heated to 300 K over 50 ps in the NVT ensemble with positional restraints on all heavy atoms using a force constant of 5 kcal/mol/Å^2. Positional restraints were gradually released over 6 ns in the NPT ensemble to fully equilibrate the systems. Production runs were performed in the isothermal-isobaric ensemble (1 atm and 300 K), employing smooth particle mesh Ewald (SPME) electrostatics, 10 Å cut-off for short-range non-bonded interactions and 2-fs time step. After 100 ns of unrestrained MD, 200 ns of accelerated molecular dynamics was employed by boosting both the total potential and the dihedral potential. Calculated values for boost parameters were 8812 kcal/mol and 155 kcal/mol, respectively. Eight snapshots leading up to the base extrusion event were selected and then replicated 10 times. The replicas were each simulated for 100 ns of free unbiased molecular dynamics, reinitializing velocities for each replica. All simulations were performed using the AMBER16 code with the AMBER Parm14SB parameter set[189].

4.6.3 Path Optimization with the Partial Nudged Elastic Band (PNEB) Method

We then carried out path optimization with the string method, based on the swarms of trajectories method, requiring definition of a lower dimensional space. Twenty-eight images from the PNEB path (including the two fixed end points) were sampled along the initial pathway defined in collective variable space. Two collective variables were defined by using an RMSD collective variable (RMSD computed over a selection of atoms relevant to the extrusion transition) and pseudo-dihedral angle denoted in Figure S6. We refined these structures using a swarm of 20 short (2-ps) simulations launched from each image. Images were updated based on mean drift in each swarm, redistributing between end states and relaxing with 980-steps unconstrained and 20-steps restrained simulations. At least 200 iterations were completed for each string. 200 ns of unconstraint simulations for each image (total 11.2 μs) were then performed by taking the final converged structure for both PNEB (28 images) and string methods (28 images). Graphics of the movie were prepared by Chimera[190].

4.6.4 MSM Construction

All TICA calculations, clustering, and MSM construction were performed using the PyEMMA software[110]. Backbone torsions of the interrogated base and the distances between base pairs and the guanidinium group of Arg275 were used as descriptive coordinates to define the TICA projections. For base interrogation, all independent components (ICs) were computed using a lag time of 50 ps (25 steps). The combined trajectories were then projected onto the first two ICs. The trajectory frames were then clustered in the projection space using the k-means algorithm, producing 800 clusters. From the clustering data, implied timescales were generated by estimating the transition probability matrix at different lag times. Using these results, 5 macrostates and a lag time of 50 ps (25 steps) were chosen to construct the MSMs for all three base interrogation systems. For the 5caC-DNA base eversion path, independent components were calculated using a lag time of 100 ps (50 steps). The combined trajectories were then projected onto the first two ICs. Trajectory frames were then clustered using uniform time clustering, a method in which data points are selected uniformly in time and assigned using a Voronoi discretization. This produced 800 clusters, from which the implied timescales were then estimated. Based on these results, 6 macrostates and a lag time of 80 ps (40 steps) was chosen for MSM construction.

CHAPTER 5. TRANSCRIPTION INITIATION MACHINERY FUNCTIONAL DYNAMICS AND GENETIC DISEASE

5.1 Abstract

Transcription pre-initiation complexes (PIC) are vital assemblies whose function underlies protein gene expression. Cryo-EM advances have begun to uncover their structural organization. Yet, functional analyses are hindered by incompletely modeled regions. Here we integrate all available cryo-EM data to build a practically complete human PIC structural model. This enables simulations that reveal the assembly's global motions, define PIC partitioning into dynamic communities and delineate how structural modules function together to remodel DNA. We identify key TFIIE/p62 interactions linking core-PIC to TFIIH. P62 rigging interlaces p34, p44 and XPD while capping XPD DNA-binding and ATP-binding sites. PIC kinks and locks substrate DNA, creating negative supercoiling within the Pol II cleft to facilitate promoter opening. Mapping Xeroderma Pigmentosum, Trichothiodystrophy, and Cockayne syndrome disease mutations onto defined communities reveals clustering into three mechanistic classes, affecting TFIIH helicase functions, protein interactions and interface dynamics.

5.2 Introduction

Complexes of RNA Polymerase II (Pol II) are foundational for transcription since all mRNA in eukaryotic cells originates from Pol II synthesis[191-194]. Additionally, Pol II transcribes most small regulatory non-coding RNAs controlling gene expression levels and acting in gene silencing. As transcription regulation governs all fundamental aspects of cell biology loss of transcriptional control is a hallmark of many autoimmune disorders, cancers, neurological, metabolic and cardiovascular diseases[195-199]. To begin transcription, Pol II depends on key general transcription factors (GTFs: TFIIA, TFIIB, TFIID, TFIIF, TFIIS, TFIIE and TFIIH) that recognize promoter DNA and assemble with the polymerase into a pre-initiation complex (PIC)[194, 199-202]. After PIC assembly, the initial closed promoter complex (CC) transitions into an open complex (OC), in which the melted single-stranded DNA template is inserted into the Pol II active site. This transient OC is then converted into an initial transcribing complex (ITC), competent to synthesize mRNA. When the nascent RNA chain grows to a critical length, Pol II clears the promoter and a stable elongation complex (EC) ensues[203, 204]. Formation of PIC and its conversion into a productive elongation complex are key for transcription regulation[205]. Yet, the molecular architecture of the PIC and its associated functional dynamics remain incompletely understood.

Importantly, with the “resolution revolution” in cryo-electron microscopy, structures of these molecular machines have recently come into view[206, 207]. Two recent cryo-EM studies achieved near atomic visualization of core Pol II PICs (excluding the mobile TFIIH GTF) in multiple states (CC, OC and ITC) and enabled side-by-side comparison of the conformational states leading to a competent elongation complex[206, 208]. Two subsequent studies showed TFIIH structure both in the absence (apo-TFIIH) and in the presence of core PIC (holo-PIC)[207, 209]. These breakthrough studies elucidated eukaryotic pre-initiation complex architectures; yet, the respective models were incomplete (>20% of residues unassigned in sequence or not modelled) and, therefore, unsuitable as starting points for detailed molecular dynamics simulations and analysis of the dynamic PIC molecular machine.

Here we synthesized all available EM data to produce the most complete atomistic model of the human PIC to date. All previously omitted/unassigned regions have now been modelled into the corresponding EM densities (Figure 5.2.1), including the ten-subunit TFIIH GTF and its flexible kinase (CAK) module. The overall assembly conformation was fitted into EM density of the CC holo-PIC. The quality of the new models makes them entirely suitable for molecular dynamics (MD) simulations on massively parallel computing platforms. Thus, we employed extensive MD simulations to unveil the functional dynamics of Pol II holo- and core-PICs. Modeling the above systems (comprised of >1,000,000 atoms) used resources of the Texas Advanced Computing Center and the Oak Ridge Leadership Computing Facility. Importantly, these analyses reveal the hierarchical organization of the PIC machinery into dynamic communities and unveil how its interwoven structural elements function together to remodel the DNA substrate and facilitate promoter opening. Strikingly, a mapping of patient-derived TFIIH mutations onto the newly discovered dynamic communities showed that mutations were clustered at critical junctures in the TFIIH dynamic network. Thus, the results provide a new level of understanding into PIC molecular mechanism and the etiology of three devastating autosomal recessive genetic disorders - Xeroderma pigmentosum (XP, cancer), trichothiodystrophy (TTD, aging) and Xeroderma pigmentosum/Cockayne syndrome (XP/CS, development, cancer). Importantly, our methods and models provide a roadmap for future structural, biochemical and mutational experiments to understand the interplay between TFIIH structural disruption and the complex XP, XP/CS and TTD disease phenotypes[210-215].

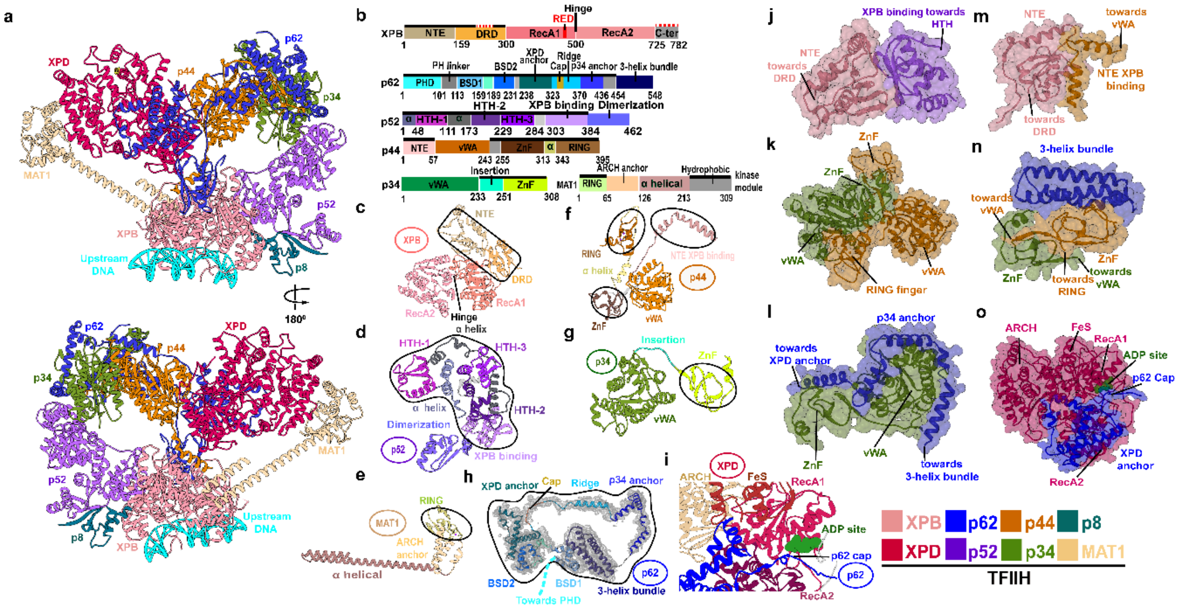

Figure 5.2.1 - TFIIH integrative structure model based on comparative analysis of cryo-EM densities reveals “molecular rigging” formed by p62 and p44. a, Anterior and posterior cartoon views of TFIIH GTF where missing regions or entire domains and proteins are built for XPB, p62, p52, p44, p34 and MAT1. Newly modeled p62 and p44 subunits act as molecular rigging, interlinking TFIIH. b, Motif schematic highlighting newly modeled regions (solid black lines). Two small regions in XPB, not present in the EM maps, were not modeled (red dashed lines). Abbreviations denote: DRD - damage recognition domain; NTE - N-terminal extension; HTH-helix-turn-helix. c-h, Cartoon of TFIIH subunits with newly modeled regions circled. h, Representative cryo-EM electron density from apo-TFIIH overlaid with p62 (PHD domain not shown). i, Zoom view of p62 cap region overlaying the XPD ATP binding cleft. Space filling views highlighting Interactions newly modeled in j, XPB NTE and p52; k, p34 and p44; l, p62 helices and p34; m, XPB N-terminus with p44; n, p62 3-helix bundle and p34 plus p44; and o, p62 and XPD.

5.3 Results

5.3.1 The molecular architecture of TFIIH underlies its role in the human PIC

The opening of promoter DNA by Pol II and the formation of the nascent transcription bubble critically depend on the transcription factor TFIIH[195, 216-219]. Specifically, a mechanism has been proposed, in which TFIIH-induced DNA translocation toward Pol II creates negative DNA supercoiling inside the polymerase cleft to facilitate promoter opening[206, 208, 220-222]. To unravel the functional role of TFIIH within the PIC, we first constructed a suitably complete model of human apo-TFIIH. Model building was based on comparative analysis of cryo-EM densities for apo-TFIIH (EMDB accession code: EMD-3802)[209] and yeast core-PIC/TFIIH/DNA (EMDB accession code: EMD-3846)[207]. To guide our initial TFIIH model into the holo-PIC cryo-EM density, we employed the cascade MDFF method[223]. TFIIH and core-PIC were separately flexibly fitted into the closed-state human holo-PIC density (EMDB accession code: EMD-3307)[206] and then combined to assemble the full PIC/TFIIH/DNA complex.

The structural model (Figure 5.2.1) resulting from the above protocol reveals newly modelled TFIIH regions that are demarcated in Figure 5.2.1b and Table 5.5.1, indicating >95% completeness. TFIIH is the most complex of all general transcription factors and encompasses ten protein subunits.34 Seven subunits form the TFIIH core (Figure 5.2.1a). Two helicase subunits, XPB and XPD, are adjacent while four intermediate subunits (p8, p52, p34, p44) lie in a characteristic horseshoe shape. The p62 subunit is the most extended: it traverses and interlaces the surfaces of p34, p44 and XPD. The MAT1 subunit connects the XPB DNA-damage recognition domain (DRD) to the XPD ARCH domain via a 86-Å long α-helix and a helical bundle (Figure 5.2.1a, b, e). The remaining three TFIIH subunits (CDK7, Cyclin-H and part of MAT1) form the flexible kinase (CAK) subcomplex, which is positioned away from the TFIIH core and is key for transcription regulation[225, 226].

Two subunits, XPB and XPD, are conceivably the most central to TFIIH's function in transcription[227, 228] and possess independent helicase and ATP-hydrolysis activities[229-232]. Yet, in transcription XPD serves a structural role and its helicase activity is suppressed. In contrast, XPB engages promoter DNA downstream of the transcription start site (TSS) (Figure 5.2.1a, Figure 5.3.1), and its ATPase activity is obligatory for Pol II initiation. Human XPB features two RecA-like lobes (denoted RecA1 and RecA2), the DRD, and an N-terminal extension domain (NTE) (Figure 5.2.1b,c). The DRD and NTE domains were built de novo in our model after tracing the entire length of XPB in the apo-TFIIH EM density. The DRD domain recognizes distortions in DNA41 and may act in DNA damage detection. Not surprisingly, it has structural similarity to the mismatch recognition domain (MRD) of MutS/MSH[233] and the SMARCAL1 HARP domain[234]. The NTE domain, important for anchoring XPB within TFIIH34, consists of three α-helices and five β-strands (residues 1-159) that contact the XPB-interacting domain of p52 (Figure 5.2.1c,d). Two human disease mutations map to the NTE, supporting its functional significance. Thus, the structural elements uncovered in our more complete TFIIH model underscore the GTF's dual functional role: (i) to unwind dsDNA and facilitate promoter opening by Pol II PIC; and (ii) to recognize damaged DNA and enable nucleotide excision repair (NER).

5.3.2 TFIIE, MAT1 and p62 are Critical for the Integrity of the Core-PIC/TFIIH Interface

In Figure 5.3.1 the TFIIH GTF is revealed in the context of the complete PIC assembly. Notably, the holo-PIC has a bipartite architecture with Pol II Rpb4/7, TFIIE, p62 and MAT1 principally responsible for the interface between core-PIC and TFIIH (Figure 5.3.2a, d and e; Table 5.5.2). Specifically, the MAT1 RING domain lodges in-between the Rpb4 and Rpb7 chains of Pol

II while also contacting α4 and α6 helices of TFIIE GTF. Additionally, MAT1 ARCH anchor domain lies between the ARCH domain of XPD and Rpb4, thus stabilizing the TFIIH/core-PIC interface. The lower half of the interface highlights the critical role of p62. We built the entire length of p62 by comparing the human and yeast EM densities. We furthermore confirmed positional assignments of all p62 domains (N-terminal plekstrin homology domain (PHD), two BSD domains - BSD1 and BSD2, XPD/p34 anchor and C-terminal 3-helix bundle) by matching our model to existing chemical cross-linking data[209, 224] (Figure 5.2.1b, h, n, l, o). Importantly, two domains from p62 (BSD2 and PHD) directly bind TFIIE through its α7 helix and adjacent loops (Figure 5.3.2e). Interfacial interactions include β-sheet formation with one strand from TFIIE and the other from p62 (e.g. p62 PHD domain forms secondary structure with the TFIIE acidic patch; Figure 5.3.2f) to strengthen the interface. Interestingly, we found that yeast and human PIC differ in the contacts at the TFIIH/core-PIC juncture. In yeast, the PHD domain extends to make direct contact with the Pol II core[207]. An analogous interaction in human PIC is impossible as the linker leading into the PHD domain is shortened by >50 residues. Instead, the PHD domain is unambiguously positioned between the XPB and XPD subunits in all existing holo-PIC EM reconstructions[206]. Crosslinking data[224] also supports this positioning, showing that PHD forms crosslinks to XPB in the human but not in the yeast PIC. The linker deletion in human PIC has more subtle effects on the lower half of the TFIIH/core-PIC interface as compared to yeast PIC. Furthermore, our model supports regulatory as well as structural roles: one p62 region (residues 274-293) tracks along the DNA path on XPD, based on a recent XPD ortholog structure[235] and another (residues 333-342) caps the XPD ATPase as suitable to influence DNA binding and ATPase function, respectively (Figure 5.2.1i and Figure 5.2.1o).

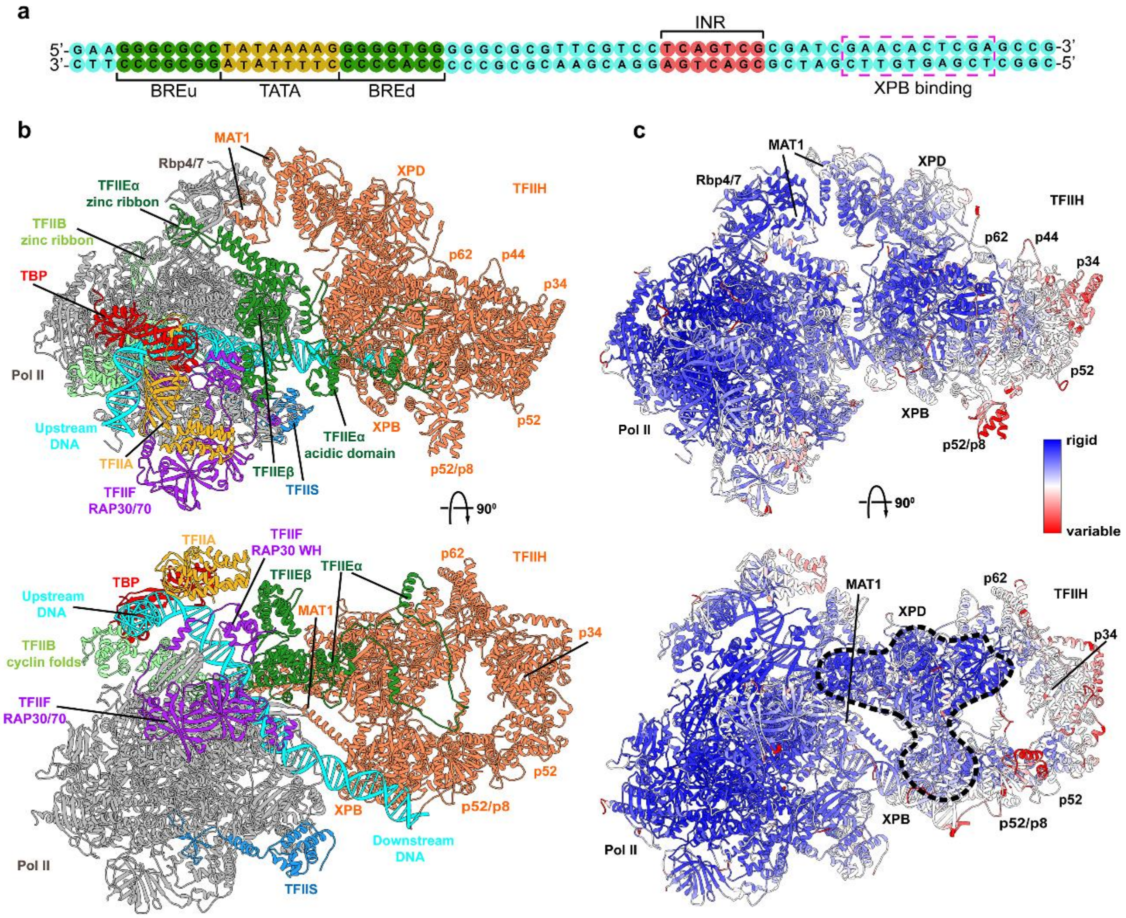

Figure 5.3.1 - TFIIE links human PIC to TFIIH. a, Sequence of the DNA substrate in PIC. b, Cartoon of human PIC structure including TFIIH highlighting non-TFIIH subunits. Most striking is TFIIE which crosses over a quarter of TFIIH. Model is based on integration and comparative analysis of cryo-EM densities for human apo-TFIIH, human closed-state holo-PIC density, and yeast core-PIC/TFIIH/DNA c, Computed B-factors mapped onto the PIC-TFIIH structural model with values colored from high (red) to low (blue) reveal a network of stable interactions. Black dashed outline highlights an unexpected rigid anchor region between TFIIH and PIC.

Notably, our results highlight a key role of TFIIE GTF (comprised of TFIIEα and TFIIEβ for PIC structural integrity as it holds a central position in the PIC assembly. TFIIEβ is a crucial constituent of the core-PIC, forming a cap over the Pol II cleft and the DNA substrate and making functionally important contacts with TFIIF. TFIIEα on the other hand, consists primarily of three α-helices (α7, α8, α9) and a β5 strand, which are splayed on the surface of TFIIH and connected by long flexible linkers (Figure 5.3.2.1 and Figure S5). Specifically, α9 binds BSD1 and the 3-helix bundle of p62; α8 interacts with the p62 PHD domain and the α7-BSD2 interaction is critically important for the TFIIH/core-PIC interface. The unusual engagement mode between TFIIEα and TFIIH highlights the key architectural role of TFIIE for assembling the pre-initiation complex. In essence, TFIIE serves as a structural adhesive to link TFIIH to the rest of the transcription initiation machinery – a finding that supports and extends current understanding of why TFIIE is required for TFIIH recruitment to the PIC[236, 237].

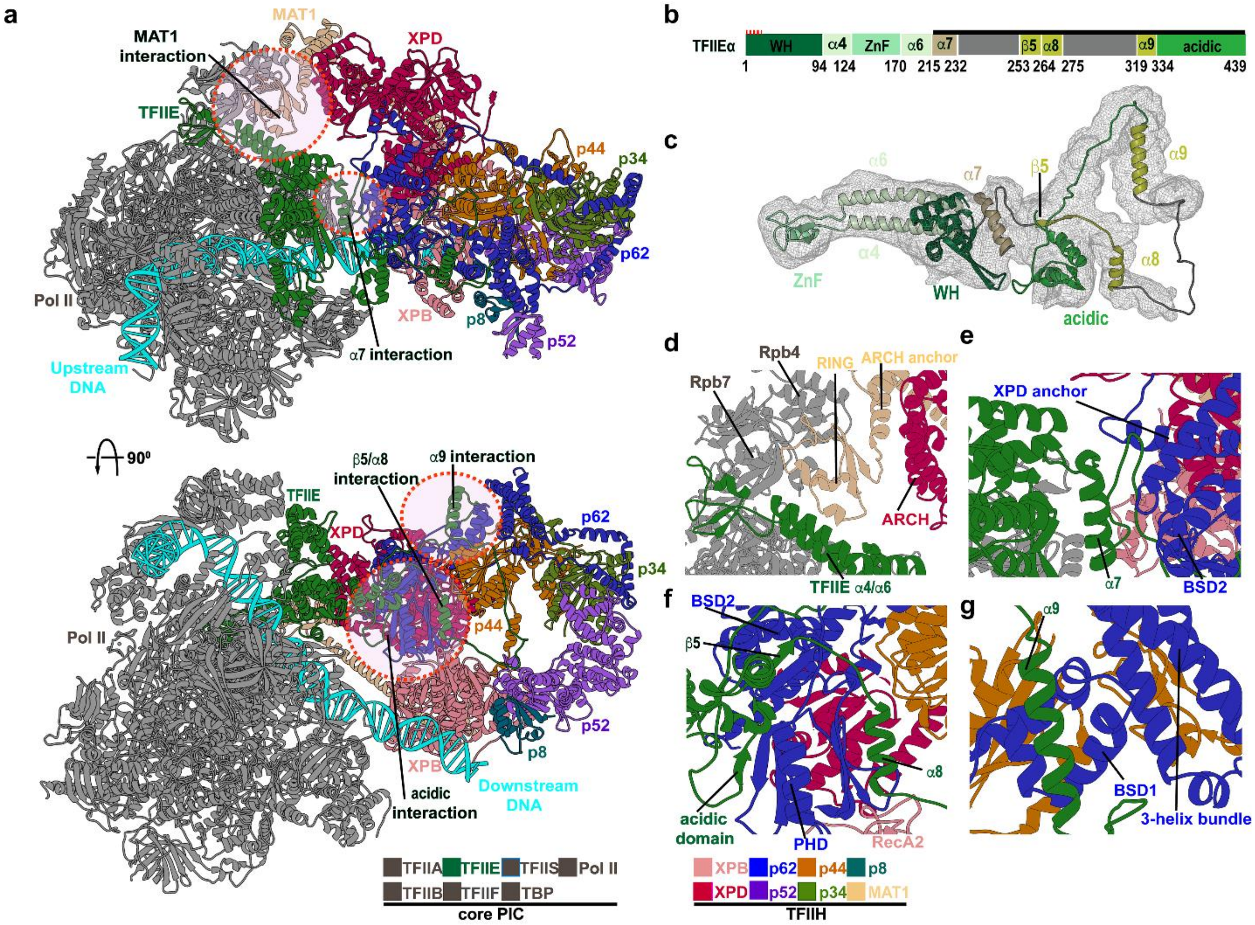

Figure 5.3.2 – TFIIE, MAT1 and p62 are critical for the integrity of the core-PIC/TFIIH interface. a, Human PIC structure in cartoon representation with colored TFIIH subunits. Circles demark zoomed regions in d.-g. b, Domain schematic of TFIIEα. c, TFIIEα cartoon. d, MAT1 - core PIC interaction. The MAT1 RING-finger docks into a groove between the Pol II stalk subunit Rpb7 and TFIIE α4/7 helices. The RING-finger connects to the ARCH anchor which binds the XPD ARCH domain. e, TFIIEα helix α7 is wedged between TFIIE winged helix (eWH) domain and p62. f, TFIIE β5/α7 and acidic domain interacts with p62 PHD and BSD2. g, TFIIE helix α9 binds p62 BSD1 domain adjacent to the p62 3-helix bundle.

5.3.3 Promoter Opening is Linked to the Global Motions and Dynamic Networks Within TFIIH

Our holo-PIC model provided an excellent starting point to initiate MD simulations aimed at addressing the role of dynamics and conformational switching in driving the multifaceted functional responses of the transcription initiation machinery. We performed ~300-ns production simulations of holo-PIC (TFIIH/core-PIC/DNA) and core-PIC. Specifically, holo-PIC is a ~1M Dalton complex, encompassing substrate DNA and some 31 individual protein chains. Addition of solvent and counterions resulted in simulation systems of >1,000,000 atoms that required supercomputing resources. To begin to dissect the staggering complexity of the simulations, we first tested if the relative rigidity or flexibility of the numerous PIC structural elements was linked to their putative functional roles. To this end, we mapped the computed B-factors from the simulation data onto the structure of holo-PIC (Figure 5.3.1b). The core of Pol II (Rpb1 and Rpb2 chains) is the most rigid scaffold within the PIC, and also the best resolved region in cryo-EM (local resolution going down to ~3 Å)[206]. The low B-factors support the importance of this region, which establishes the path of the DNA substrate and confines it within the Pol II cleft. The DNA duplex upstream of the initiation region (INR; Figure 5.3.1a) is also structurally rigid, especially in the TBP-associated TATA box region. The downstream portion of DNA is more mobile, and its mobility is linked to the motion of XPB. Notably, there is a ridge of stability extending across the TFIIH/core-PIC interface starting with the Pol II core, continuing through TFIIE and encompassing core XPD, portion of p62 and lobe1 of XPB. In contrast, the middle domains (p8, p52, p34 and p44) are dynamic and appear to participate in concerted global motions.

To analyze and visualize such global PIC motions, we relied on two methods - principal component analysis (PCA)[238] and community network analysis[72]. Briefly, PCA is a dimensionality reduction technique that involves three steps: 1) computing the matrix of residue-residue covariances from the MD trajectories; 2) diagonalizing the covariance matrix to yield eigenvectors (principal modes) and eigenvalues (mean square fluctuations); and 3) projecting the trajectory onto the principal modes to yield principal components. The first few principal modes are especially important as they capture the slow, large-amplitude motions that are also the most functionally significant. Our focus was on the second and third principal modes of PIC (denoted PC2 and PC3). The PC2 mode reflects an out-of-plane twisting movement of TFIIH with respect to the Pol II core above the ringed plane of TFIIH defined by the p44, p34, p52 and XPD and XPB subunits. Interestingly, PC3 represents the in-plane swing motion of TFIIH that could push the DNA substrate toward the Pol II cleft, leading to DNA bending and deformation. Although differing in detail, both PC2 and PC3 imply the DNA substrate is rigidly held by Pol II in the TBP region while the downstream DNA duplex is held and pushed by XPB whose motion is directed by rotational movement of the TFIIH lever arm (comprised of p44, p34, p52 and p8).

We employed community network analysis to partition PIC into dynamic communities (tightly connected clusters of residues that move together as modules). In this approach, the PIC assembly was mapped onto a graph wherein each protein residue is a node and edges connect nodes that are in contact. All edges were assigned weights based on the covariance matrix data from the MD simulation. The Girvan-Newman algorithm was then used to subdivide the graph into strongly connected components. The magnitude of allosteric communication between communities was then quantified by estimating the total betweenness for all edges that connect individual communities (betweenness is defined as the number of shortest paths that cross an edge). We identified sixteen dynamic communities in TFIIH, which were color-coded and mapped onto the holo-PIC structure (Figure 5.3.3a) and graphed the level of dynamic communication between communities (Figure 5.3.3b). An important observation from this analysis was that the motions of the two XPB lobes are largely decoupled. Lobe1 (community L) carries much stronger connection to community A (largely comprised of XPD). Lobe2 (community C) appears more closely associated with communities O and J that form the tip of the TFIIH lever arm (subunits p52, p34, p44) and community H that includes part of p62. In PC2 and PC3 lobe2 and p62 fragment from community H move concertedly in the same direction whereas lobe1 exhibits smaller magnitude motions and is most closely correlated to XPD (Figure 5.3.3c, g). Community network analysis also nicely captures the fact that the motion of XPB lobe1 is coupled to the motion of MAT1 through the long helix (strong connection between communities L and N). The XPD ARCH domain separates into its own community (Figure 5.3.3d,h) while the TFIIE (community E) is in communication with p62 (community H) and the motions of these elements occur in the same direction (Figure 5.3.3e, i). Figure 5.3.3f and j capture the directionality and concerted rotational motion of the TFIIH lever arm. Notably, communities I and B and communities B and J are both separated by hinge regions.

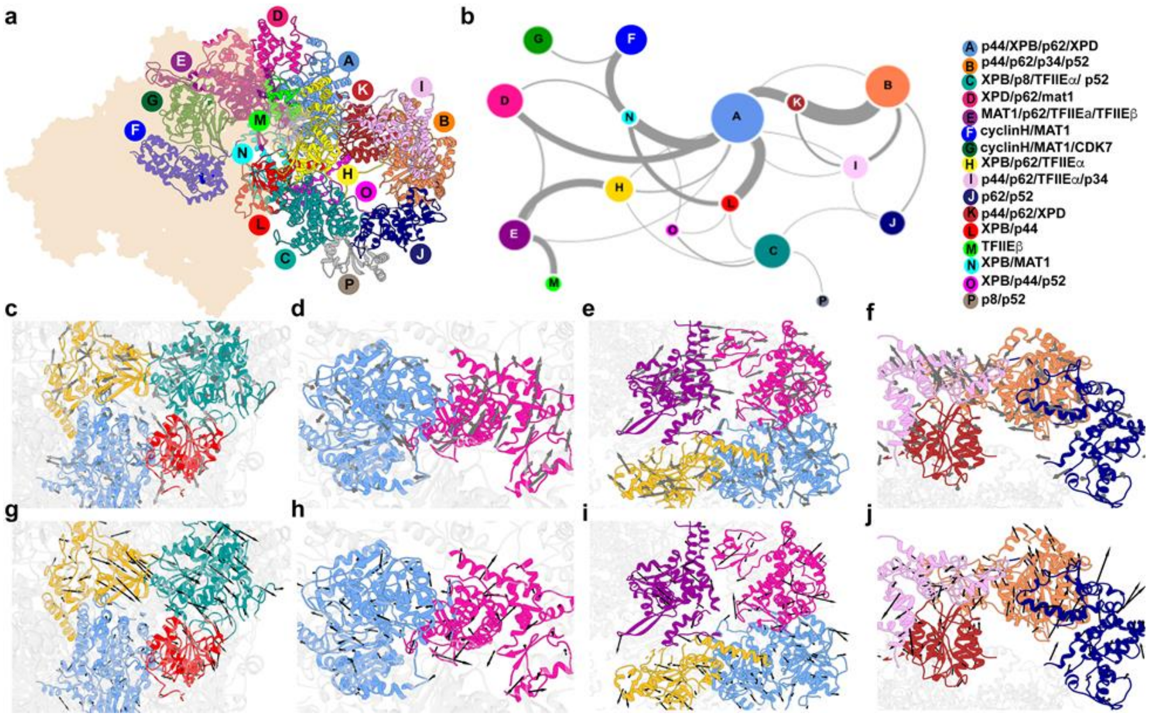

Figure 5.3.3 - Community networks underlying TFIIH functional dynamics. a, Communities identified from dynamic network analysis that transcend simple subunit divisions. b, Graph of allosteric communication among communities. Colored by community, nodes are sized by number of residues in each community. Thickness of edges between community pairs correlate to magnitude of dynamic communication (betweenness). c-j, Directional community motions (arrows) and magnitude (arrow size) of the corresponding component of the eigenvector for c, A, C, H and L along the second principal component; d. A and D along the second principal component; e, A, D, E and H along the second principal component; f, B, I, J and K along the second principal component; g, A, C, H and L along the third principal component; h, A and D along the third principal component; i, A, D, E and H along the third principal component; and j, B, I, J and K along the third principal component.

5.3.4 The Global Dynamics of PIC Facilitates Substrate DNA Bending and Deformation

To examine the effect of global PIC dynamics on the DNA substrate, we analyzed the DNA substrate path through the Pol II cleft for our holo-PIC and core-PIC simulations (Figure 5.3.4). The DNA substrate traverses the Pol II cleft undergoing an ~ 90° bend at the position of TBP. The DNA path continues relatively straight between the Pol II Rpb2 and Rpb1 subunits. Interactions with the Rpb5 subunit lead to ~135° kinking of the DNA near the transcription start site. Surprisingly, the DNA duplex is kinked in the INR region with and without TFIIH. Figure 5b shows a 2-D histogram of the MD data in terms of two angles ϕ and θ representing the bending and twisting of the DNA duplex around the axis defined by the straight region preceding the INR. While the bending angle θ appears to be approximately the same for holo- and core-PIC, the orientation angle ϕ is different, spanning a far greater range for core-PIC. Thus, besides kinking the DNA substrate, TFIIH also locks it into a fixed orientation.

The kinked DNA conformation could be attributed entirely to interactions with structural elements from the Pol II cleft (Rpb2, Rpb1 coiled-coil and clamp head, Rpb5) (Figure 5.3.4c and d). XPB also induces a slight bend in the DNA as it passes between the two lobes but does not affect the INR region. Overlaying canonical A-DNA onto the INR region shows that the DNA substrate is not only bent but significantly under-wound (Figure 5.3.4e,f). Negative DNA supercoiling should facilitate promoter opening. Thus, we propose that the role of TFIIH may be to further unwind the DNA until base flipping occurs leading to the formation of a nascent transcription bubble. Consistent with this proposition, negative DNA supercoiling relieves the requirement for TFIIH in basal transcription at multiple promoters[239, 240].

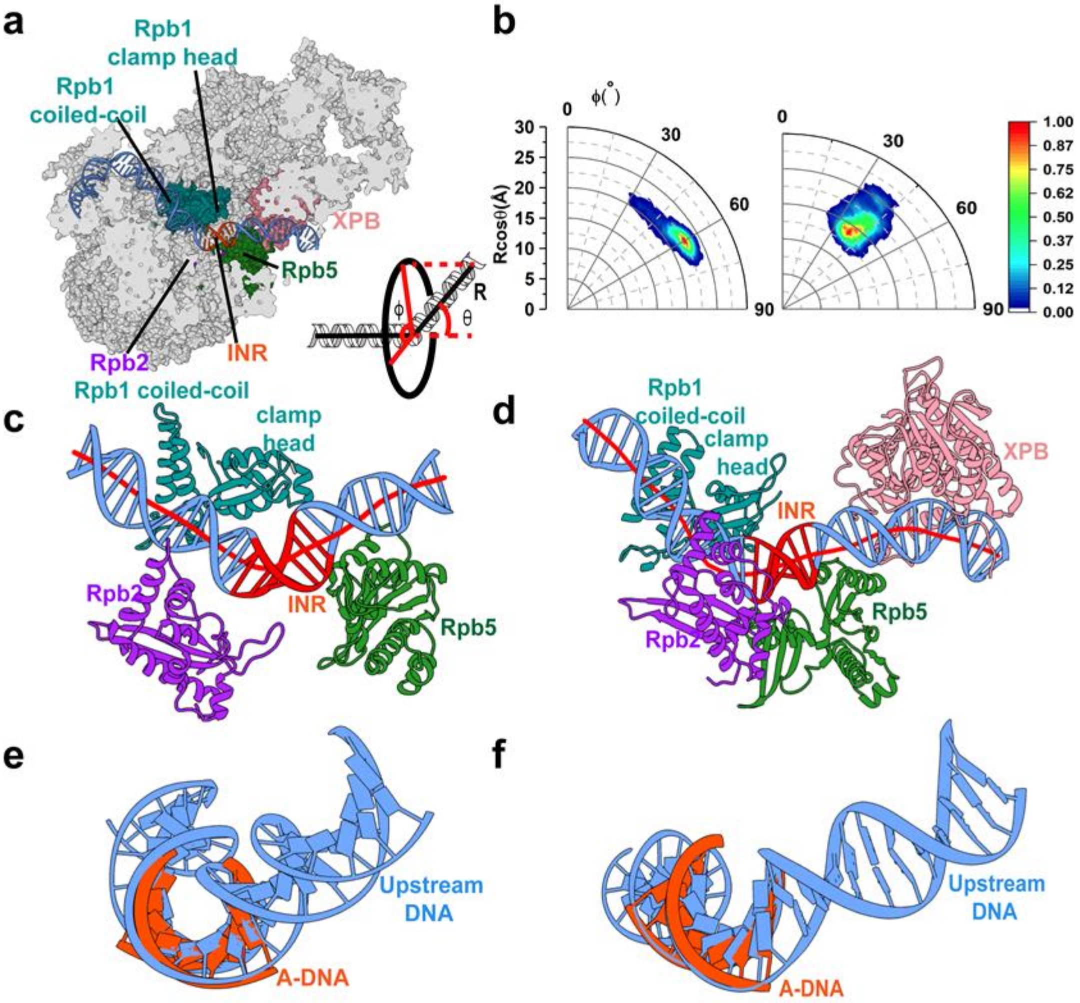

Figure 5.3.4 - Pol II induces DNA bending and distortion neighboring the initiation site. a, DNA path within the PIC. The inset defines two geometric variables orientation angle (ϕ) and bending angle (θ) used in the analysis of the MD trajectories. b, Histogram of DNA bending and orientation angles in polar coordinates from the MD simulations of core PIC (left) and holo PIC (right). c, d, Pol II structural elements induce bending of downstream DNA from simulations of core PIC (c) and holo PIC (d). DNA axes (red lines) are computed by the CURVES+ code. e, f, Pol II induces structural distortion in the INR region besides bending. Comparison with canonical A DNA shows that in the INR region the DNA duplex is significantly underwound in core PIC and holo-PIC.

5.3.5 Disease Mutations Cluster at Critical Junctures of the TFIIH Dynamic Network

XP, TTD and XP/CS are three distinct autosomal recessive genetic disorders. Patients are generally compound heterozygotes carrying two different mutations – one on each allele. The expressed phenotype results from contributions of both alleles[241]. In general, TFIIH disease-causing point mutations relate specific sequence sites to distinct pathways and phenotypes: XP mutants are NER defective, TTD mutants have partial transcription defects, XP/TTD mutations exhibit both defects, and XP/CS mutations exhibit defective global genome repair (GGR) and transcription coupled repair (TCR)[211-215, 242].

To link molecular features at these sites to distinct pathways and disease phenotypes, we mapped the 36 single-site patient mutations (Figure 5.3.5 and Table S3) onto the dynamic PIC model defined here. These fall within the XPD, XPB and p8 subunits of TFIIH and the WH2 domain of TFIIEβ[243] (Figure 5.3.5b), largely coinciding with the anchor region of reduced flexibility between TFIIH and PIC identified in our dynamics study (Figure 5.3.1b). Strikingly 80% of disease mutations localize to XPD helicase domain with none in transcription-essential XPB helicase domains. Two are in the XPB N-terminus (defined in our structure); one in XPD Fe-S domain, two in XPD Arch domain, and one in p8. Notably, 20 XPD mutants localize to RecA2 and the 8 in RecA1 cluster close to RecA2, pointing to RecA2's pivotal role in TFIIH repair function. In our structural model, RecA2 is the central and most interactive helicase domain: it connects XPB, p62, and p44. The XPD helix (residues 712-725) at the three-community junction is a hot spot for patient disease mutations. Intriguingly, many XPD mutations lie along the path of p62 as it traverses across XPD, suggesting that p62 has regulatory as well as scaffolding functions: one p62 region (residues 274-293) tracks the XPD DNA binding groove and another (residues 333-342) caps the XPD ATP site. Most patient mutations map to secondary structure ends or loops, highlighting their significance (Table S3). Whereas half the TTD mutations fall within helices, this position is rare for XP and XP/CS mutations.

As has been noted[212, 229], the single-site disease mutations fall across TFIIH sequence and structure in an irregular spatial pattern (Figure 5.3.5). Yet, our largely complete TFIIH and dynamics model unveils TTD, XP, XP/TTD and XP/CS mutations in the context of the full TFIIH machinery, unhampered by missing regions in XPB, p44, and p62. We find that patient mutations cluster primarily at protein and community interfaces (70%). The largest cluster at the intersection of communities A (XPD) and K (XPD, p44, p62), demarks the XPD/p44/p62 interaction as functionally important. Three mutations are directly at the interface: R592P (TTD), R722W (XP/TTD) and A725P (TTD, XP/CS/TTD), and 12 more are immediately adjacent: Y18H (XP/CS, TTD), G47R(XP/CS), S51F (XP/TTD); L485P (XP), R487G(TTD), R616P/Q (TTD), D673G (TTD), G675R(XP/CS), A594P(TTD), A596P(TTD), A717G (XP), and G713R (TTD). Switches to glycine and proline, which have the greatest impact on local backbone flexibility and dynamics, dominate in this region unveiling the critical functional role of dynamics at the XPD/p44 junction. The other key interfaces include communities A and D with mutations R636W (TTD), R112H/C(XP/TTD), A and L with mutations D681N (XP), R683W/Q (XP, XP/TTD), and R511Q (XP). Unlike XPD, neither of the XPB mutations (F99S (XP/CS) or T119P (TTD)) are in the helicase domains, but they center in four communities: C (XPB, p8, p52, TFIIEα), L (XPB, p44), N (XPB, Mat1) and O (XPB, p44, p52). Similarly, the sole p8 mutation, L21P (TTD), is at a community interface between C (XPB, p8, p52, TFIIEα) and P (p8, p52). Importantly, all these clusters correspond to critical junctures in the TFIIH dynamic network.

Considering mutations by disease, we find that 14 TTD or XP/TTD mutations mapped to protein-protein interfaces or in helices at protein interfaces in three of the TFIIH subunits (p8, XPB and XPD) - an insight enabled by our fully modeled TFIIH complex. The TTD mutations map to all helicase interfaces: XPB with XPD, p8, p44, or p52; or XPD with Mat1, p44, and p62. We therefore propose that TTD mutations disrupt protein-protein interfaces (directly or through breaks in helices at interfaces) to weaken assembly of TFIIH subunits while retaining residual XPB helicase activity for essential transcription function. Notably, TTD mutations were previously shown to result in lower levels of unstable TFIIH[210, 215]. Recently, two TTD patient-derived mutations, A150P and D187Y, have been discovered in TFIIE.54 In our model, these positions map to the WH2 domain of TFIIEβ and near the linchpin TFIIE α7 helix, which is key for the integrity of the TFIIE/p62 interface. These mutations are positioned to reduce WH2 stability by disrupting secondary structure packing (Table S3). In turn, this is expected to weaken interactions with TFIIEα and the interface between dynamic communities E (TFIIE) and H (p62).

All XP mutations map to XPD near to its exterior and fall into three categories. One set (R112H/C, R511H/C, S541R, Y542C, R601L/W, and R683W/Q) tracks the proposed DNA path on XPD, also traced by p62. A second set (residues S51F, T76A, D234N, and C663R) neighbors the Walker A, B motifs and are expected to reduce ATPase activity. These two sets substitute charged or polar for bulky hydrophobic residues, potentially disrupting XPD-DNA binding and ATPase activities during NER. The third set (L485P, D681N, R683W/Q, A717G, and R722W) is on the opposite side where they appear to weaken interfaces with XPB, p44, and p62, suggesting that these XPD interactions with other TFIIH subunits are required for NER function. Mutation in residues 683 and 722 also have a TDD phenotype in some patients, suggesting that for these mutations both assembly and NER are defective, consistent with our structure.

All but one XPD XP/CS mutations lie close to p62, which is split between communities A and H. p62 wraps the XPD core such that five XP/CS mutants (Y18H, G47R, G602D, R666W, G675W/Q) are near p62, which spans several dynamic communities, running along the XPD DNA binding path, capping ATPase site and linking TFIIH to TFIIE and the core-PIC. The one XPB XP/CS mutation, F99S, is near p44, another rigging-like protein that links XPB to XPD, p62 and p34. XP/CS mutations are primarily at the center of communities and typically increase rigidity or distort conformation by removing glycines or changing to more rigid side chains (Table S3). Like a broken gear in a machine, these changes appear to break down community coordination. Therefore, we propose XP/CS mutations weaken TFIIH dynamic coordination explaining its hallmark TCR defects[242].

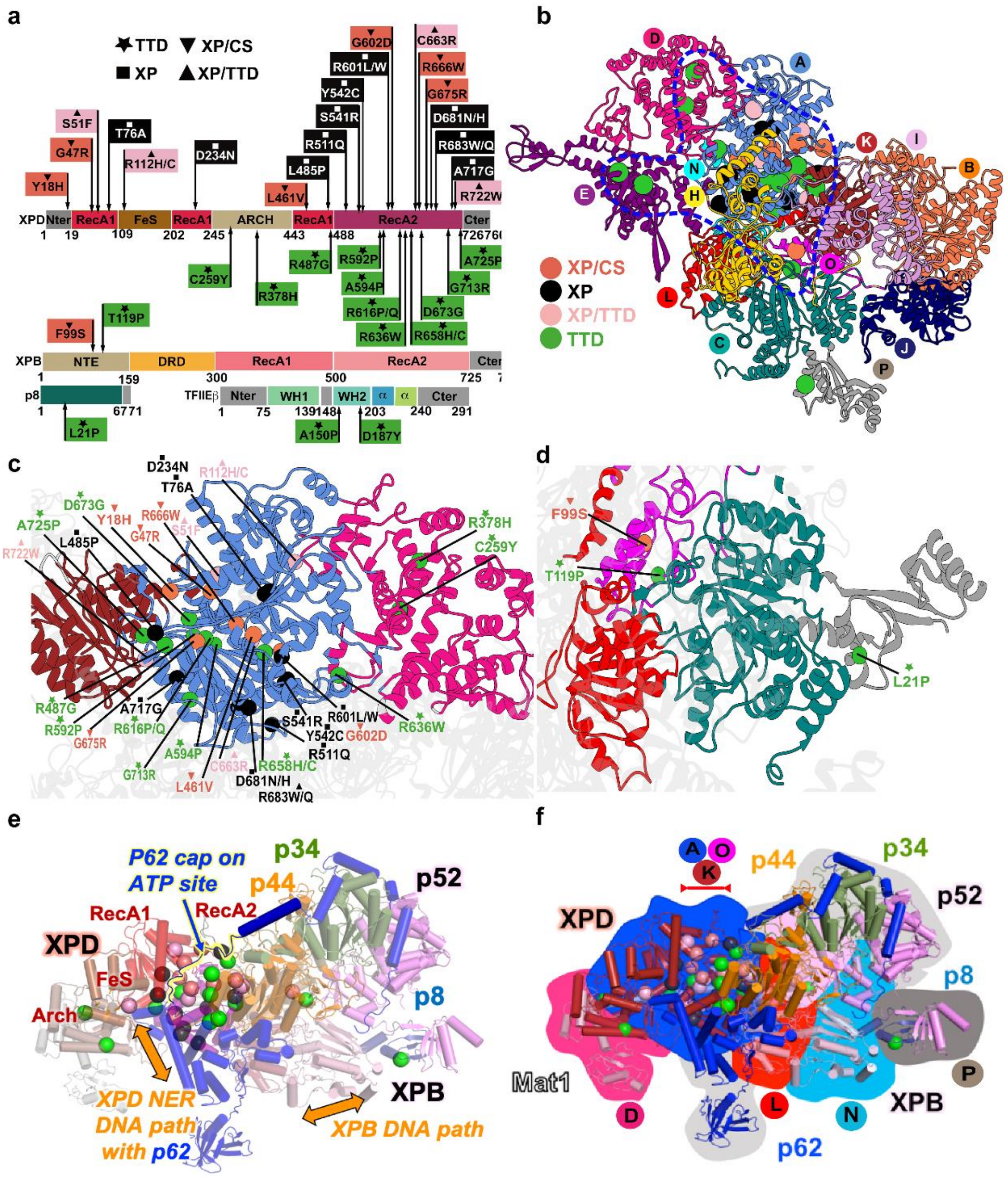

Figure 5.3.5 - Human disease mutations mapped onto TFIIH and TFIIE show distinct patterns within protein-protein and community interfaces. a, TTD, XP/TTD, XP and XP/CS point mutations mapped onto XPD, XPB, and TFIIE protein schematic do not co-localize by disease on primary sequence. b, Map of human disease mutations (spheres) onto TFIIH structure as cartoon colored by community show biased localization (blue outline). c, Zoom view of mutations on XPD. d, Zoom view of XPB and p8 mutations. e, Mutations and function mapped onto TFIIH cartoon colored by subunit. Regions of p62 and p44 are removed for clarity. f, Overlay of disease mutations, protein chain (cartoon view), and communities (background color). Community K is above the plane of the page of Community A, as demarked. View matches e.

5.4 Discussion

Transcription initiation complexes are amazingly dynamic macromolecular machines whose function and regulation underlie all of gene expression. By synthesizing cryo-EM data, we built and analyzed the functional dynamics of a practically complete atomic model of human PIC. Our results support a model for promoter opening in which the TFIIH XPB subunit serves as a DNA translocase to bend and unwind the DNA duplex in the cleft of Pol II. In this model, Pol II by itself can induce structural deformation in the INR region of the DNA template. TFIIH's role is to lock the downstream DNA duplex in a well-defined orientation and to use a ratcheting mechanism to induce further negative supercoiling in the INR region. This action of TFIIH leads to strain-induced base flipping and the formation of a nascent transcription bubble. For this model to be operational the DNA duplex must be rigidly locked upstream of the transcription start site. The 90° bend in the DNA induced by TBP serves precisely this purpose. A second requirement is for the molecular motor twisting the downstream DNA duplex to be firmly attached to the rest of the initiation machinery. Consistent with this notion, we find a ridge of structural stability extending from the Pol II core through the TFIIE/p62 interface and into the central XPD subunit of TFIIH while also encompassing lobe1 of XPB. Disruption of this ridge by point mutations impairs TFIIH function as seen by the striking clustering of patient-derived mutants in this region.

Finally, DNA remodeling to produce the transcription bubble appears to be a global conformational transition that critically depends on TFIIH global dynamics. Our MD simulation powerfully elucidates concerted motions of this complex machinery. Specifically, we show that the two lobes of the XPB translocase move independently. Lobe1's motion is primarily correlated with the XPD subunit while lobe2 tracks the large-scale collective motion of the TFIIH lever arm (subunits p44, p34, p52). In the absence of ATP such motion is bidirectional. However, during cycles of ATP binding and hydrolysis by XPB the backward motion could be disallowed, leading to the simultaneous unwinding and pushing of the DNA toward the TSS. Interestingly, mapping of patient-derived mutations onto the TFIIH community structure further informs the above model for promoter opening. Specifically, this initiation model and disease phenotypes argue that mutants affecting XPD stability and/or its community integrity are functionally significant. Conversely, interfaces between the p44, p34 and p52 subunits lack disease causing mutations, indicating that either mutations at these interfaces are so disruptive as to be invariably lethal or, more probably, point mutations in the TFIIH lever arm are too distal from XPB to disrupt the global motions and, thus, result in mild or no disease phenotype.

Overall, our results elucidate the functional dynamics of the human transcription initiation machinery. This knowledge enabled structural and mechanistic insights into preinitiation complex assembly, promoter recognition, DNA melting, transcription regulation and the roles of general transcription factors therein. Collectively, these methods and results provide a framework for future experiments aimed at unraveling the intricate molecular choreography of TFIIH in nucleotide excision repair as well as transcription initiation.

5.5 Methods

5.5.1 Building the apo-TFIIH Model

To create a model of apo-TFIIH, we used two existing cryo-EM densities: apo human TFIIH (EMDB accession code: EMD-3802)[209] and yeast PIC (EMDB accession code: EMD-3846)[207]. The corresponding structure[209] (PDB accession codes: 5OF4) served as a starting point for model building. The following TFIIH elements had no known structural homologues and were, therefore, built de novo: the XPB DNA-damage recognition domain (DRD) (residues 159-300), the XPB N-terminal extension domain (NTE) (residues 1-158), the p52 XPB binding domain (residues 284-384), the p34 insertion (residues 233-251), the p44 N-terminus and α-helix insertion (residues 1-57 and 313-343), the MAT1 ARCH anchor and helices (residues 65-309), the p62 subunit except the BSD1 and PHD domain (residues 101-173 and 159-548). We used the GeneSilico metaserver[244] for consensus secondary structure prediction and applied the results to establish the sequence register in the EM density. Individual secondary-structure fragments were constructed using COOT[245] to generate a backbone only model by tracing the EM density. The resulting polypeptide chain segments were connected by extending the main-chain trace. Side chain orientations were built and manually inspected/corrected based on the electron density. Bulky residues such as phenylalanine, tyrosine, tryptophan, and arginine were instrumental in validating model construction and sequence registration.

To model the p62 BSD1 domain, the NMR structure of the p62 BSD1 domain (PDB ID: 2DII) was rigid-body docked into the EM density. The human p52 subunit resembles the yeast Tfb2 and shares 40% sequence identity (64% similarity). Therefore, the p52 helix-turn-helix (HTH) domain (residues 1-282) was constructed by homology modeling using MODELER 9V15 software[246] and alignment to yeast Tfb2 (PDB ID: 5OQJ)[207]. To model the p44 subunit and the p34 ZINC finger domain (ZnF), the structures of the yeast Ssl1 and Tfb4 subunit (PDB ID: 5OQJ) [207] were used as templates to construct the human p34/p44 ZnF homology structure. The RING domain of p44 was taken from the X-ray p34/p44 structure (PDB accession code: 5O85 [247]) and docked into the density after overlaying the p34 vWA domain. To model MAT1, the solution structure of the human MAT1 RING domain (PDB ID:1G25)[248] was docked into the density ascribed to MAT1 RING by superposing the yeast Tfb1 density onto the human MAT1 density. We then built the entire apo TFIIH structure by docking the newly constructed XPB, p62, p52, p34, p44 and MAT1 subunits into the apo-TFIIH EM density.

5.5.2 Building the holo-PIC Model

To model TFIIH holo-PIC (core-PIC/TFIIH/DNA), we first docked the human apo-TFIIH structure into the yeast PIC density (accession code: EMD-3846)[207]. We then used the cascade molecular dynamics flexible fitting (cMDFF) method[223] to fit apo-TFIIH into the density allowing the model to be fit sequentially to a series of maps (computationally blurred derivatives of the original map with lower-resolution). Thus, larger-scale features of the Gaussian-smoothed EM densities guided the initial stages of flexible fitting. Smaller-scale refinements were then introduced when fitting to the higher-resolution maps. The p62 PHD domain was excluded from fitting to the yeast PIC density as its orientation in the human PIC density was clearly different. Gaussian-smoothed maps were generated using Chimera[190] starting with a half-width of $\sigma = 3$ Å and decreasing by 1 Å for each subsequent map. In total, 4 maps were used in cMDFF runs, including the original EM density. The structure was independently fitted using direct MDFF[128] to each individual map obtained by Gaussian blurs. 4-ns MDFF simulations were performed at each of the 4 resolutions to achieve convergence during the cMDFF protocol. The final structure from cMDFF was further refined with direct MDFF to the human PIC density (accession code: EMD-3307)[209]. The MDFF bias was applied in each stage with a scaling factor ξ of 0.2.

5.5.3 Building the Interface of core-PIC with TFIIH

To model the C-terminus of TFIIEα, we employed the human closed-complex PIC EM density and the yeast PIC EM density (EMDB accession codes: EMD-3307[206] and EMD-3846[207], respectively). The C-terminal region of TFIIEα (residues 215-439) comprises three helices and one beta-strand (α7/α5/α8/α9) connected by loop regions. We inspected all holo-PIC densities (EMD-3307, EMD-8132 and EMD-8133)[206] and positioned α7 between the TFIIEα WH domain and the p62 BSD2 domain. The predicted α5/α8/α9 elements were modeled based on the corresponding positions in the yeast PIC density (EMD-3846). The NMR structure for a short TFIIEα C-terminal segment bound to PHD (PDB accession code: 2RNR)[236] was positioned between XPD and XPB and subsequently validated by crossing-linking[224]. We then built the linker connecting the p62 BSD1 and PHD domains. TFIIH, TFIIEα C-terminus, the p62 PHD domain, core-PIC (PDB accession code: 5IYA)[206] and duplex DNA were then fitted to the human closed-complex PIC density (EMD-3307) to produce the complete holo-PIC assembly. The apo-TFIIH and holo-PIC models were refined in real space with the PHENIX package[249, 250]. MolProbity results for the apo-TFIIH and holo-PIC models are presented in Table 5.5.3. Map-to-model cross correlation values of 0.75 and 0.72 were computed for apo-TFIIH and holo-PIC, respectively. Table 5.5.4 summarizes map-to-model validation statistics for TFIIH fitted against the EMD-3802 and EMD-3846 cryo-EM maps.

5.5.4 Molecular Dynamics Simulations of core-PIC and holo-PIC

To address the functional dynamics of the holo-PIC and core-PIC assemblies, we performed extensive molecular dynamics simulations. The systems were set up with the TLeap module of AMBER 14[131] and solvated with TIP3P water molecules[29]. The minimum distance from the surface atoms of the complex to the edge of the periodic simulation box was 12.0 Å. In addition to Na+ counterions to neutralize the total charge in the simulation box, we introduced 150-mM NaCl concentration to mimic physiological conditions. Energy minimization was conducted for 3000 steps with fixed protein backbone atoms and for an additional 1500 steps with harmonic restraints on the backbone atoms (k = 10 kcal mol^{-1} Å^{-2}). The temperature of the simulated systems was then gradually increased to 300 K over 500 ps of dynamics in the NVT ensemble. The equilibration was continued for another 4 ns in the NPT ensemble, and the harmonic restraints were gradually released. Production runs were performed in the NPT ensemble (1 atm and 300 K) for 300 ns for each of the two models of the core PIC and holo-PIC. The particle mesh Ewald (PME) method was used to treat long-range electrostatic interactions. The r-RESPA multiple time step method[25] was employed with a 2-fs time step for bonded, 2-fs time step for short-range nonbonded interactions, and 4-fs for long-range electrostatic interactions. The short-range nonbonded interactions were evaluated by using a cutoff distance of 10 Å and a switching function at 8.5 Å. All covalent bonds to hydrogen atoms were constrained using the SHAKE algorithm. The simulations were carried out with the NAMD 2.12 code[251, 127] and the AMBER Parm14SB force field[126]. Snapshots from the MD trajectories were collected at 2.0 ps intervals. We then selected and sampled 50,000 conformations from the last 280 ns of the MD trajectories for principal component analysis (PCA) and community network analysis. DNA structural parameters were analyzed with the program CURVES+[171]. All figures were generated using UCSF Chimera[190].

5.5.5 Principal Component Analysis

Principal component analysis (PCA) was performed based on the covariance matrix whose elements are defined as:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$$

where x_i is a Cartesian coordinate of atom i, and $\langle x_i \rangle$ represents the time average over all the configurations obtained in the simulation. In PCA, diagonalization of the covariance matrix yields a complete set of orthogonal eigenvectors with corresponding eigenvalues. Eigenvectors with the larger eigenvalues contribute more to the total variance in the data and, therefore, to the overall motion seen in the MD trajectories. In this way, PCA helped to separate out the slower global motions, essential for PIC dynamics. Prior to construction of the covariance matrix the MD trajectory was aligned on a reference configuration to remove all translational and rotational motion. The covariance matrix *C* was computed using all protein Cα atoms and P atoms in the DNA backbone and then diagonalized to produce the PCA eigenvectors and eigenvalues. PCA analysis was performed using the CPPTRAJ module in AmberTools17[252].

5.5.6 Community Network Analysis

The dynamic community network of TFIIH was constructed using the NetworkView plugin in VMD[79, 253]. In community network analysis, the protein topology was represented as a collection of nodes connected by edges whose weights were derived from the MD simulation. Nodes were associated with the protein Cα atoms. Edges were added to the network connecting pairs of in-contact nodes. Two non-adjacent nodes were connected by an edge if the nodes are within 4.5 Ǻ of each other for at least 75% of the simulation trajectory. The edge weights, wi,j, were computed from the correlation coefficients, $c_{i,j}$, of the i-j node pair:

$$w_{i,j} = -ln(|c_{i,j}|)$$

Here, $c_{i,j}$ is the residue-residue correlation calculated between the i-j residue pair in the MD trajectory. Residue-residue correlations were calculated using the program CARMA[254]. The contact map was generated within the NetworkView plugin. After constructing the TFIIH network the Girvan-Newman algorithm was employed to determine the underlying community structure using the betweenness centrality measure[78]. The betweenness centrality measure (betweenness) of an edge is measured by calculating the number of shortest paths that cross that edge and is indicative of the probability of information transfer between nodes (protein residues). In Girvan-Newman, the betweenness is calculated for all edges and the edge with the largest betweenness value (most central edge) is subsequently removed. This process was repeated and a modularity score tracked to identify the division that resulted in an optimal community structure. The algorithm was run iteratively resulting in a modularity score of 0.871 and a network partitioning of 16 distinct communities. We then computed the summation of the betweenness for all edges between communities to determine the strength of communication between dynamically correlated sets of residues within TFIIH.

Table 5.5.1 – Summary of TFIIH and PIC structural elements and original sources used for modeling.

Protein	Chain	Size (aa)	Modeled Residues	Alternative names	Structures (PDB IDs) used for hybrid modeling
XPD	0	760	11-742	ERCC2	Residues 11-742 modeled from 5OF4
p62	1	548	1-546	GTF2H1	Residues 159-546 built de novo; p62 BSD1 domain (residues 110-158) modeled from NMR structure (2DII); p62 PHD domain (residues 1-109) modeled from NMR structure (2RNR)
p52	2	462	6-458	GTF2H4	Residues 284-384 built de novo; Residues 6-383 constructed by homology modeling using the yeast Tfb2 (5OQJ) as a template
MAT1	3	309	1-309	MAAT1	Residues 65-309 built de novo; MAT1 RING domain (residues 1-64) modeled from NMR structure (1G25)
p34	4	308	6-300	GTF2H3	Residues 233-251 built de novo; ZINC finger domain (residues 252-300) constructed by homology modeling using the yeast Tfb2 (5OQJ) as a template; vWA domain (residues 6-251) modeled from 5OF4
p8	5	71	2-67	GTF2H5	Residues 2-67 modeled from 5OF4
p44	6	395	10-394	GTF2H2	Residues 1-57, 313-343 built de novo; vWA domain (residues 58-312) modeled from 5OF5; ZINC finger domain (residues 344-394) constructed by homology modeling using the yeast Ssl1 (5OQJ) as a template
XPB	7	782	30-201 267-728	ERCC3	Residues 30-201 and 267-300 built de novo; Residues 301-728 modeled from 5OF5
CDK7	8	346	13-311	MO15	Residues 13-311 constructed by homology modelling using human CDK2 (1JSU) as a template
Cyclin H	9	323	11-286	CCNH	Residues 11-286 modeled from human Cyclin H (1KXU)
TFIIE α	Q	439	10-439	GTF2E1	Residues 10-215 modeled from human TFIIE (5GPY); Residues 215-234 built de novo; Residues 335-439 modeled from NMR structure 2RNR
TFIIE β	R	292	75-242	GTF2E2	Residues 75-242 modeled from NMR structure (2RNR)
Core PIC (pol II, TBP, TFIIA, TFIIB, TFIIF, TFIIS and DNA)					Core PIC modeled from the EM structure (5IY6)

Table 5.5.2 – Summary of interfaces between TFIIH and core-PIC

Interface	*Area (Å^2)* [1]	*N_{inter}* [2]	*N_{HB}* [3]	*N_{SB}* [4]
MAT1 – Rpb4/7	584.3	17	11	9
MAT1 – TFIIEα	320.4	13	5	5
CDK7 – Rpb4	456.7	15	3	2
TFIIEα – p62	4607.2	105	75	41
TFIIEα – TFIIH	6107.6	141	97	56
TFIIEα – core-PIC	1257.0	43	16	8
TFIIEβ – core-PIC	628.7	20	9	6

1 – Buried surface area at the interface
2 – Number of interfacial residues
3 – Number of interfacial hydrogen bonds
4 – Number of interfacial salt bridges

Table 5.5.3 – MolProbity results of apo TFIIH and holo PIC

Validation	Apo TFIIH	Holo PIC
MolProbity score	2.66	2.12
MolProbity Clashscore	23.5	7.98
Rotamers outliers (%)	1.34	0.39
Cβ deviations (%)	0.23	0.04
Ramachandran favored (%)	81.34	83.72
Ramachandran allowed (%)	17.17	13.33
Ramachandran outliers (%)	1.50	2.95

Table 5.5.4 – Human and yeast TFIIH MDFF map-to-model statistics

	MolProbity	iFSC1 (Å)[2]	iFSC2 (Å)[2]	d_Model (Å)	d_FSC_Model (Å)[3]
hTFIIH – h1	2.66	3.76	3.68	6.7	3.6/4.4/7.3
hTFIIH – h2	2.66	3.83	3.76	6.8	3.7/4.4/7.3
hTFIIH – f	2.66	4.50	4.29	4.2	3.9/4.3/7.2
yTFIIH – h1	2.46	2.82	2.80	6.7	3.9/6.4/8.7
yTFIIH – h2	2.46	2.77	2.75	6.8	4.0/6.2/8.6
yTFIIH – f	2.46	3.33	3.29	6.8	4.7/5.0/8.5

1 – All values are reported for both half maps and full map denoted h1 (half map 1), h2 (half map 2) and f (full map). hTFIIH denotes human TFIIH and yTFIIH denotes yeast TFIIH maps, respectively.

2 – Integrated FSCs (iFSC) between 12 – 4.4 Å for human TFIIH (hTFIIH) and 12 – 4.7 Å for yeast TFIIH (yTFIIH).

3 – Values reported at 0, 0.143 and 0.5 Fourier shell coefficients (FSC).

CHAPTER 6. PROSPECTIVE

The works presented in this manuscript employ numerous computational methods to address molecular mechanisms involved in maintaining genomic integrity. Sophisticated path optimization strategies coupled with extensive MD sampling proved crucial in elucidating these mechanisms. Until recently, path optimization on large systems was computational prohibitive. However, this constraint has been alleviated through recent advances in computer and GPU architecture. Perhaps more important is the evolution of structure elucidating techniques like cryo-EM and X-ray crystallography, which have also benefited from computer advancements. Indeed, the success of both path optimization and MD hinge on coordinates derived from structural biology. Thus, the two disciplines are intricately intertwined.

It is important to note that while the studies presented here elucidate new details surrounding specific biological processes, outstanding issues remain unresolved. For example, multiple substrates exist for TDG, yet, it is not entirely clear if the details of the base-flipping mechanism will remain the same for each substrate. In the case of Pol III, proofreading for the bacterial replicase is likely to be different from that of the eukaryotic replicase. Even more interesting would be determining how proofreading and replication are conducted in the context of the entire replisome. For the PIC, molecular details on the transitions between the CC, OC and ITC functional states remain elusive.

Looking forward, these issues are poised to take advantage of both structural and advanced computational methods. For TDG and the PIC, structural coordinates are already available through X-ray crystallography and cryo-EM. Additionally, the eukaryotic structures of both the lagging strand and leading strand polymerases (Pol δ and Pol ε) have been solved via cryo-EM. Perhaps even more exciting is the recent breakthrough with Pol ε, which is presented in the context of CMG helicase. Thus, these systems, and the questions they pose, are primed for computational investigation. Specifically, path optimization with PNEB could be combined with extensive sampling and Markov modeling to define the biomolecular mechanisms, while also identifying important intermediates that lie beyond the reach of conventional experimental methods. Moreover, the predictive power of these models will drive the formulation of new hypotheses, ones that are directly testable through experiment. In the end, a multi-disciplinary effort aimed at providing a complete description of these mechanisms will be key to broadening our understanding on the complex relationship between genomic machinery and disease.

REFERENCES

1. McHenry, C.S., 2011. DNA replicases from a bacterial perspective. Annual review of biochemistry, 80, pp.403-436.
2. Johansson, E. and Dixon, N., 2013. Replicative DNA polymerases. Cold Spring Harbor perspectives in biology, 5(6), p.a012799.
3. Steitz, T.A. and Steitz, J.A., 1993. A general two-metal-ion mechanism for catalytic RNA. Proceedings of the National Academy of Sciences, 90(14), pp.6498-6502.
4. Mazouzi, A., Velimezi, G. and Loizou, J.I., 2014. DNA replication stress: causes, resolution and disease. Experimental cell research, 329(1), pp.85-93.
5. Patel, P.H. and Loeb, L.A., 2001. Getting a grip on how DNA polymerases function. Nature structural biology, 8(8), pp.656-659.
6. Kunkel, T.A., 2004. DNA replication fidelity. Journal of Biological Chemistry, 279(17), pp.16895-16898.
7. Lindahl, T., 1993. Instability and decay of the primary structure of DNA. nature, 362(6422), pp.709-715.
8. Wallace, S.S., 2014. Base excision repair: a critical player in many games. DNA repair, 19, pp.14-26.
9. Krokan, H.E. and Bjørås, M., 2013. Base excision repair. Cold Spring Harbor perspectives in biology, 5(4), p.a012583.
10. Kim, Y.J. and M Wilson III, D., 2012. Overview of base excision repair biochemistry. Current molecular pharmacology, 5(1), pp.3-13.

11. Slupphaug, G., Mol, C.D., Kavli, B., Arvai, A.S., Krokan, H.E. and Tainer, J.A., 1996. A nucleotide-flipping mechanism from the structure of human uracil–DNA glycosylase bound to DNA. Nature, 384(6604), pp.87-92.

12. Friedman, J.I. and Stivers, J.T., 2010. Detection of damaged DNA bases by DNA glycosylase enzymes. Biochemistry, 49(24), pp.4957-4967.

13. Wiebauer, K. and Jiricny, J., 1990. Mismatch-specific thymine DNA glycosylase and DNA polymerase beta mediate the correction of GT mispairs in nuclear extracts from human cells. Proceedings of the National Academy of Sciences, 87(15), pp.5842-5845.

14. Hendrich, B., Hardeland, U., Ng, H.H., Jiricny, J. and Bird, A., 1999. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature, 401(6750), pp.301-304.

15. Horst, J.P. and Fritz, H.J., 1996. Counteracting the mutagenic effect of hydrolytic deamination of DNA 5-methylcytosine residues at high temperature: DNA mismatch N-glycosylase Mig. Mth of the thermophilic archaeon Methanobacterium thermoautotrophicum THF. The EMBO journal, 15(19), pp.5459-5469.

16. Thomas, M.C. and Chiang, C.M., 2006. The general transcription machinery and general cofactors. Critical reviews in biochemistry and molecular biology, 41(3), pp.105-178.

17. Buratowski, S., Hahn, S., Guarente, L. and Sharp, P.A., 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. Cell, 56(4), pp.549-561.

18. Cheung, A. C. M., Sainsbury, S. & Cramer, P. Structural basis of initial RNA polymerase II transcription. EMBO J. 30, 4755–4763 (2011).

19. Liu, X., Bushnell, D.A., Silva, D.A., Huang, X. and Kornberg, R.D., 2011. Initiation complex structure and promoter proofreading. Science, 333(6042), pp.633-637.

20. Leach, A.R. and Leach, A.R., 2001. *Molecular modelling: principles and applications*. Pearson education.

21. Bayly, C.I., Cieplak, P., Cornell, W. and Kollman, P.A., 1993. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, *97*(40), pp.10269-10280.

22. Ponder, J.W. and Case, D.A., 2003. Force fields for protein simulations. In *Advances in protein chemistry* (Vol. 66, pp. 27-85). Academic Press.

23. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. and Berendsen, H.J., 2005. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, *26*(16), pp.1701-1718.

24. MacKerell Jr, A.D., Banavali, N. and Foloppe, N., 2000. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers: Original Research on Biomolecules*, *56*(4), pp.257-265.

25. Tuckerman, M.B.B.J.M., Berne, B.J. and Martyna, G.J., 1992. Reversible multiple time scale molecular dynamics. *The Journal of chemical physics*, *97*(3), pp.1990-2001.

26. Darden, T., York, D. and Pedersen, L., 1993. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics*, *98*(12), pp.10089-10092.

27. Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H. and Pedersen, L.G., 1995. A smooth particle mesh Ewald method. *The Journal of chemical physics*, *103*(19), pp.8577-8593.

28. Roux, B. and Simonson, T., 1999. Implicit solvent models. *Biophysical chemistry*, *78*(1-2), pp.1-20.

29. Harrach, M.F. and Drossel, B., 2014. Structure and dynamics of TIP3P, TIP4P, and TIP5P water near smooth and atomistic walls of different hydroaffinity. *The Journal of Chemical Physics*, *140*(17), p.174501.
30. Lippert, R.A., Predescu, C., Ierardi, D.J., Mackenzie, K.M., Eastwood, M.P., Dror, R.O. and Shaw, D.E., 2013. Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *The Journal of chemical physics*, *139*(16), p.10B621_1.
31. Litniewski, M., 1993. Molecular dynamics method for simulating the constant temperature volume and temperature-pressure system. *The Journal of Physical Chemistry*, *97*(15), pp.3842-3848.
32. Toxvaerd, S., 1993. Molecular dynamics at constant temperature and pressure. *Physical Review E*, *47*(1), p.343.
33. Sugita, Y. and Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, *314*(1-2), pp.141-151.
34. Grubmüller, H., 1995. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, *52*(3), p.2893.
35. Hénin, J. and Chipot, C., 2004. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *The Journal of chemical physics*, *121*(7), pp.2904-2914.
36. Voter, A.F., 1997. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Physical Review Letters*, *78*(20), p.3908.
37. Hamelberg, D., Mongan, J. and McCammon, J.A., 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics*, *120*(24), pp.11919-11929.

38. Steiner, M.M., Genilloud, P.A. and Wilkins, J.W., 1998. Simple bias potential for boosting molecular dynamics with the hyperdynamics scheme. *Physical Review B*, *57*(17), p.10236.
39. Rahman, J.A. and Tully, J.C., 2002. Puddle-skimming: An efficient sampling of multidimensional configuration space. *The Journal of chemical physics*, *116*(20), pp.8750-8760.
40. Miao, Y., Sinko, W., Pierce, L., Bucher, D., Walker, R.C. and McCammon, J.A., 2014. Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. *Journal of Chemical Theory and Computation*, *10*(7), pp.2677-2689.
41. Torrie, G.M. and Valleau, J.P., 1974. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chemical Physics Letters*, *28*(4), pp.578-581.
42. Torrie, G.M. and Valleau, J.P., 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, *23*(2), pp.187-199.
43. Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H. and Kollman, P.A., 1992. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of computational chemistry*, *13*(8), pp.1011-1021.
44. Souaille, M. and Roux, B., 2001. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer physics communications*, *135*(1), pp.40-57.
45. Grossfield, A., 2013. WHAM: the weighted histogram analysis method, version 2.0. 9. *Available at membrane. urmc. rochester. edu/content/wham. Accessed November*, *15*, p.2013.

46. Fischer, S. and Karplus, M., 1992. Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chemical physics letters*, *194*(3), pp.252-261.

47. Bolhuis, P.G., Chandler, D., Dellago, C. and Geissler, P.L., 2002. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry*, *53*(1), pp.291-318.

48. Weinan, E., Ren, W. and Vanden-Eijnden, E., 2002. String method for the study of rare events. *Physical Review B, 66*(5), p.052301.

49. Pan, A.C., Sezer, D. and Roux, B., 2008. Finding transition pathways using the string method with swarms of trajectories. *The journal of physical chemistry B, 112*(11), pp.3432-3440.

50. Maragliano, L., Fischer, A., Vanden-Eijnden, E. and Ciccotti, G., 2006. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *The Journal of chemical physics*, *125*(2), p.024106.

51. Bergonzo, C., Campbell, A.J., Walker, R.C. and Simmerling, C., 2009. A partial nudged elastic band implementation for use with large or explicitly solvated systems. *International journal of quantum chemistry*, *109*(15), pp.3781-3790.

52. Jonsson, H., Mills, G. and Jacobsen, K.W., 1998. Classical and quantum dynamics in condensed phase systems.

53. Elber, R. and Karplus, M., 1987. A method for determining reaction paths in large molecules: Application to myoglobin. *Chemical Physics Letters*, *139*(5), pp.375-380.

54. Frauenfelder, H., Sligar, S.G. and Wolynes, P.G., 1991. The energy landscapes and motions of proteins. *Science, 254*(5038), pp.1598-1603.

55. Noé, F., Krachtus, D., Smith, J.C. and Fischer, S., 2006. Transition networks for the comprehensive characterization of complex conformational change in proteins. *Journal of chemical theory and computation*, *2*(3), pp.840-857.

56. Krivov, S.V. and Karplus, M., 2004. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proceedings of the National Academy of Sciences*, *101*(41), pp.14766-14770.

57. Muff, S. and Caflisch, A., 2008. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics*, *70*(4), pp.1185-1195.

58. Prinz, J.H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J.D., Schütte, C. and Noé, F., 2011. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, *134*(17), p.174105.

59. Jolliffe, I.T., 2002. Principal components in regression analysis. *Principal component analysis*, pp.167-198.

60. Molgedey, L. and Schuster, H.G., 1994. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters*, *72*(23), p.3634.

61. Naritomi, Y. and Fuchigami, S., 2011. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *The Journal of chemical physics*, *134*(6), p.02B617.

62. Schwantes, C.R. and Pande, V.S., 2013. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of chemical theory and computation*, *9*(4), pp.2000-2009.

63. Chodera, J.D., Singhal, N., Pande, V.S., Dill, K.A. and Swope, W.C., 2007. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of chemical physics*, *126*(15), p.04B616.

64. Fischer, I. and Poland, J., 2005, January. Amplifying the block matrix structure for spectral clustering. In *Proceedings of the 14th annual machine learning conference of Belgium and the Netherlands* (pp. 21-28). Citeseer.

65. Weber, M., 2003. Improved Perron cluster analysis.

66. Weber, M., 2013. Adaptive spectral clustering in molecular simulation. In *Classification and Data Mining* (pp. 147-154). Springer, Berlin, Heidelberg.

67. Röblitz, S. and Weber, M., 2013. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Advances in Data Analysis and Classification*, *7*(2), pp.147-179.

68. Kannan, R., Vempala, S. and Vetta, A., 2004. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, *51*(3), pp.497-515.

69. Kube, S. and Weber, M., 2007. A coarse graining method for the identification of transition rates between molecular conformations. *The Journal of chemical physics*, *126*(2), p.024103.

70. Metzner, P., Schütte, C. and Vanden-Eijnden, E., 2009. Transition path theory for Markov jump processes. *Multiscale Modeling & Simulation*, *7*(3), pp.1192-1219.

71. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. and Weikl, T.R., 2009. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, *106*(45), pp.19011-19016.

72. Wu, H., Paul, F., Wehmeyer, C. and Noé, F., 2016. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proceedings of the National Academy of Sciences*, *113*(23), pp.E3221-E3230.

73. Wu, H. and Noé, F., 2014. Optimal estimation of free energies and stationary densities from multiple biased simulations. *Multiscale Modeling & Simulation*, *12*(1), pp.25-54.

74. Mey, A.S., Wu, H. and Noé, F., 2014. xTRAM: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states. *Physical Review X*, *4*(4), p.041018.

75. Shirts, M.R. and Chodera, J.D., 2008. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, *129*(12), p.124105.

76. Newman, M.E., 2003. The structure and function of complex networks. *SIAM review*, *45*(2), pp.167-256.

77. Girvan, M. and Newman, M.E., 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), pp.7821-7826.

78. Newman, M.E., 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), pp.8577-8582.

79. Eargle, J. and Luthey-Schulten, Z., 2012. NetworkView: 3D display and analysis of protein· RNA interaction networks. *Bioinformatics*, *28*(22), pp.3000-3001.

80. Benkovic, S.J., Valentine, A.M. and Salinas, F., 2001. Replisome-mediated DNA replication. *Annual review of biochemistry*, *70*(1), pp.181-208.

81. Joyce, C.M. and Benkovic, S.J., 2004. DNA polymerase fidelity: kinetics, structure, and checkpoints. *Biochemistry*, *43*(45), pp.14317-14324.

82. Johnson, K.A., 2010. The kinetic and chemical mechanism of high-fidelity DNA polymerases. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1804*(5), pp.1041-1048.

83. Steitz, T.A., 1998. A mechanism for all polymerases. *Nature*, *391*(6664), pp.231-232.

84. Steitz, T.A., 1999. DNA polymerases: structural diversity and common mechanisms. *Journal of Biological Chemistry*, *274*(25), pp.17395-17398.

85. Purohit, V., Grindley, N.D. and Joyce, C.M., 2003. Use of 2-aminopurine fluorescence to examine conformational changes during nucleotide incorporation by DNA polymerase I (Klenow fragment). *Biochemistry*, *42*(34), pp.10200-10211.

86. Fernandez-Leiro, R., Conrad, J., Yang, J.C., Freund, S.M., Scheres, S.H. and Lamers, M.H., 2017. Self-correcting mismatches during high-fidelity DNA replication. *Nature structural & molecular biology*, *24*(2), pp.140-143.

87. Fernandez-Leiro, R., Conrad, J., Scheres, S.H. and Lamers, M.H., 2015. cryo-EM structures of the E. coli replicative DNA polymerase reveal its dynamic interactions with the DNA sliding clamp, exonuclease and τ. *Elife*, *4*, p.e11134.

88. Johnson, A. and O'Donnell, M., 2005. Cellular DNA replicases: components and dynamics at the replication fork. *Annu. Rev. Biochem.*, *74*, pp.283-315.

89. Lamers, M.H., Georgescu, R.E., Lee, S.G., O'Donnell, M. and Kuriyan, J., 2006. Crystal structure of the catalytic α subunit of E. coli replicative DNA polymerase III. *Cell*, *126*(5), pp.881-892.

90. Lamers, M.H. and O'Donnell, M., 2008. A consensus view of DNA binding by the C family of replicative DNA polymerases. *Proceedings of the National Academy of Sciences*, *105*(52), pp.20565-20566.

91. Johansson, E. and Dixon, N., 2013. Replicative DNA polymerases. *Cold Spring Harbor Perspectives in Biology*, *5*(6), p.a012799.

92. Rock, J.M., Lang, U.F., Chase, M.R., Ford, C.B., Gerrick, E.R., Gawande, R., Coscolla, M., Gagneux, S., Fortune, S.M. and Lamers, M.H., 2015. DNA replication fidelity in Mycobacterium tuberculosis is mediated by an ancestral prokaryotic proofreader. *Nature genetics*, *47*(6), pp.677-681.

93. Barros, T., Guenther, J., Kelch, B., Anaya, J., Prabhakar, A., O'Donnell, M., Kuriyan, J. and Lamers, M.H., 2013. A structural role for the PHP domain in E. coli DNA polymerase III. *BMC structural biology*, *13*(1), pp.1-12.

94. Beese, L.S. and Steitz, T.A., 1991. Structural basis for the 3′-5′ exonuclease activity of Escherichia coli DNA polymerase I: a two metal ion mechanism. *The EMBO journal*, *10*(1), pp.25-33.

95. Scheuermann, R., Tam, S., Burgers, P.M., Lu, C. and Echols, H., 1983. Identification of the epsilon-subunit of Escherichia coli DNA polymerase III holoenzyme as the dnaQ gene product: a fidelity subunit for DNA replication. *Proceedings of the National Academy of Sciences*, *80*(23), pp.7085-7089.

96. Studwell-Vaughan, P.S. and O'Donnell, M., 1993. DNA polymerase III accessory proteins. V. Theta encoded by holE. *Journal of Biological Chemistry*, *268*(16), pp.11785-11791.

97. Taft-Benz, S.A. and Schaaper, R.M., 2004. The θ subunit of Escherichia coli DNA polymerase III: a role in stabilizing the ε proofreading subunit. *Journal of bacteriology*, *186*(9), pp.2774-2780.

98. Hamdan, S., Bulloch, E.M., Thompson, P.R., Beck, J.L., Yang, J.Y., Crowther, J.A., Lilley, P.E., Carr, P.D., Ollis, D.L., Brown, S.E. and Dixon, N.E., 2002. Hydrolysis of the 5 ‘-p-

Nitrophenyl Ester of TMP by the Proofreading Exonuclease (ε) Subunit of Escherichia coli DNA Polymerase III. *Biochemistry*, *41*(16), pp.5266-5275.

99. Georgescu, R.E., Kim, S.S., Yurieva, O., Kuriyan, J., Kong, X.P. and O'Donnell, M., 2008. Structure of a sliding clamp on DNA. *Cell*, *132*(1), pp.43-54.

100. Wang, L., Xu, X., Kumar, R., Maiti, B., Liu, C.T., Ivanov, I., Lee, T.H. and Benkovic, S.J., 2013. Probing DNA clamps with single-molecule force spectroscopy. *Nucleic acids research*, *41*(16), pp.7804-7814.

101. McInerney, P., Johnson, A., Katz, F. and O'Donnell, M., 2007. Characterization of a triple DNA polymerase replisome. *Molecular cell*, *27*(4), pp.527-538.

102. Mok, M. and Marians, K.J., 1987. The Escherichia coli preprimosome and DNA B helicase can form replication forks that move at the same rate. *Journal of Biological Chemistry*, *262*(34), pp.16644-16654.

103. Bloom, L.B., Chen, X., Fygenson, D.K., Turner, J., O’Donnell, M. and Goodman, M.F., 1997. Fidelity Of Escherichia Coli DNA Polymerase III Holoenzyme The Effects Of Β, Γ Complex Processivity Proteins And ε Proofreading Exonuclease On Nucleotide Misincorporation Efficiencies. *Journal of Biological Chemistry*, *272*(44), pp.27919-27930.

104. Ghoreishi, D., Cerutti, D.S., Fallon, Z., Simmerling, C. and Roitberg, A.E., 2019. Fast Implementation of the Nudged Elastic Band Method in AMBER. *Journal of chemical theory and computation*, *15*(8), pp.4699-4707.

105. Pan, A.C., Sezer, D. and Roux, B., 2008. Finding transition pathways using the string method with swarms of trajectories. *The journal of physical chemistry B*, *112*(11), pp.3432-3440.

106. Dodd, T., Yan, C., Kossmann, B.R., Martin, K. and Ivanov, I., 2018. Uncovering universal rules governing the selectivity of the archetypal DNA glycosylase TDG. *Proceedings of the National Academy of Sciences*, *115*(23), pp.5974-5979.

107. Paul, F., Wehmeyer, C., Abualrous, E.T., Wu, H., Crabtree, M.D., Schöneberg, J., Clarke, J., Freund, C., Weikl, T.R. and Noé, F., 2017. Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nature communications*, *8*(1), pp.1-10.

108. Shamoo, Y. and Steitz, T.A., 1999. Building a replisome from interacting pieces: sliding clamp complexed to a peptide from DNA polymerase and a polymerase editing complex. *Cell*, *99*(2), pp.155-166.

109. Beese, L.S., Derbyshire, V. and Steitz, T.A., 1993. Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science*, *260*(5106), pp.352-355.

110. Scherer, M.K., Trendelkamp-Schroer, B., Paul, F., Pérez-Hernández, G., Hoffmann, M., Plattner, N., Wehmeyer, C., Prinz, J.H. and Noé, F., 2015. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *Journal of chemical theory and computation*, *11*(11), pp.5525-5542.

111. Prinz, J.H., Keller, B. and Noé, F., 2011. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Physical Chemistry Chemical Physics*, *13*(38), pp.16912-16927.

112. Bowman, G.R., Beauchamp, K.A., Boxer, G. and Pande, V.S., 2009. Progress and challenges in the automated construction of Markov state models for full protein systems. *The Journal of chemical physics*, *131*(12), p.124101.

113. Jergic, S., Horan, N.P., Elshenawy, M.M., Mason, C.E., Urathamakul, T., Ozawa, K., Robinson, A., Goudsmits, J.M., Wang, Y., Pan, X. and Beck, J.L., 2013. A direct proofreader–clamp interaction stabilizes the Pol III replicase in the polymerization mode. *The EMBO journal*, *32*(9), pp.1322-1333.

114. Park, J., Jergic, S., Jeon, Y., Cho, W.K., Lee, R., Dixon, N.E. and Lee, J.B., 2018. Dynamics of Proofreading by the E. coli Pol III Replicase. *Cell chemical biology*, *25*(1), pp.57-66.

115. Hamdan, S., Carr, P.D., Brown, S.E., Ollis, D.L. and Dixon, N.E., 2002. Structural basis for proofreading during replication of the Escherichia coli chromosome. *Structure*, *10*(4), pp.535-546.

116. Sethi, A., Eargle, J., Black, A.A. and Luthey-Schulten, Z., 2009. Dynamical networks in tRNA: protein complexes. *Proceedings of the National Academy of Sciences*, *106*(16), pp.6620-6625.

117. Van Wart, A.T., Durrant, J., Votapka, L. and Amaro, R.E., 2014. Weighted implementation of suboptimal paths (WISP): an optimized algorithm and tool for dynamical network analysis. *Journal of chemical theory and computation*, *10*(2), pp.511-517.

118. Gahlon, H.L., Walker, A.R., Cisneros, G.A., Lamers, M.H. and Rueda, D.S., 2018. Reduced structural flexibility for an exonuclease deficient DNA polymerase III mutant. *Physical Chemistry Chemical Physics*, *20*(42), pp.26892-26902.

119. Xu, X., Yan, C., Kossmann, B.R. and Ivanov, I., 2016. Secondary interaction interfaces with PCNA control conformational switching of DNA polymerase PolB from polymerization to editing. *The Journal of Physical Chemistry B*, *120*(33), pp.8379-8388.

120. Naufer, M.N., Murison, D.A., Rouzina, I., Beuning, P.J. and Williams, M.C., 2017. Single-molecule mechanochemical characterization of E. coli pol III core catalytic activity. *Protein Science*, *26*(7), pp.1413-1426.

121. Franklin, M.C., Wang, J. and Steitz, T.A., 2001. Structure of the replicating complex of a pol α family DNA polymerase. *Cell*, *105*(5), pp.657-667.

122. Ibarra, B., Chemla, Y.R., Plyasunov, S., Smith, S.B., Lazaro, J.M., Salas, M. and Bustamante, C., 2009. Proofreading dynamics of a processive DNA polymerase. *The EMBO journal*, *28*(18), pp.2794-2802.

123. Berdis, A.J., 2009. Mechanisms of DNA polymerases. *Chemical reviews*, *109*(7), pp.2862-2879.

124. Gouge, J., Ralec, C., Henneke, G. and Delarue, M., 2012. Molecular recognition of canonical and deaminated bases by P. abyssi family B DNA polymerase. *Journal of molecular biology*, *423*(3), pp.315-336.

125. Fiser, A., Do, R.K.G. and Šali, A., 2000. Modeling of loops in protein structures. *Protein science*, *9*(9), pp.1753-1773.

126. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. and Simmerling, C., 2015. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation*, *11*(8), pp.3696-3713.

127. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K., 2005. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, *26*(16), pp.1781-1802.

128. Trabuco, L.G., Villa, E., Schreiner, E., Harrison, C.B. and Schulten, K., 2009. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods*, *49*(2), pp.174-180.

129. Gotz, A.W., Williamson, M.J., Xu, D., Poole, D., Le Grand, S. and Walker, R.C., 2012. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born. *Journal of chemical theory and computation*, *8*(5), pp.1542-1555.

130. Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S. and Walker, R.C., 2013. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of chemical theory and computation*, *9*(9), pp.3878-3888.

131. Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr, K.M., Onufriev, A., Simmerling, C., Wang, B. and Woods, R.J., 2005. The Amber biomolecular simulation programs. *Journal of computational chemistry*, *26*(16), pp.1668-1688.

132. Vanden-Eijnden, E., 2010. Transition-path theory and path-finding algorithms for the study of rare events. *Annual review of physical chemistry*, *61*, pp.391-420.

133. Politis, D.N. and Romano, J.P., 1994. The stationary bootstrap. *Journal of the American Statistical association*, *89*(428), pp.1303-1313.

134. Sheridan, R., Fieldhouse, R.J., Hayat, S., Sun, Y., Antipin, Y., Yang, L., Hopf, T., Marks, D.S. and Sander, C., 2015. Evfold. org: Evolutionary couplings and protein 3d structure prediction. *biorxiv*, p.021022.

135. Zharkov, D.O., 2008. Base excision DNA repair. *Cellular and molecular life sciences*, *65*(10), pp.1544-1565.

136. Bellacosa A & Drohat AC (2015) Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites. *DNA repair* 32:33-42.

137. Liu MY, DeNizio JE, Schutsky EK, & Kohli RM (2016) The expanding scope and impact of epigenetic cytosine modifications. *Curr Opin Chem Biol* 33:67-73.

138. Breiling A & Lyko F (2015) Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin* 8:24.

139. Kohli RM & Zhang Y (2013) TET enzymes, TDG and the dynamics of DNA demethylation. Nature 502:472-479.

140. Eden S, Hashimshony T, Keshet I, Cedar H, & Thorne AW (1998) DNA methylation models histone acetylation. *Nature* 394:842.

141. Jones PA & Laird PW (1999) Cancer epigenetics comes of age. *Nat Genet* 21:163-167.

142. Baylin SB & Jones PA (2011) A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* 11:726-734.

143. Issa JP (2004) CpG island methylator phenotype in cancer. *Nat Rev Cancer* 4:988-993.

144. Issa JP (2014) Aging and epigenetic drift: a vicious cycle. *J Clin Invest* 124:24-29.

145. Ito S*, et al.* (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 333:1300-1303.

146. Pastor WA, Aravind L, & Rao A (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat Rev Mol Cell Biol* 14:341-356.

147. Maiti A & Drohat AC (2011) Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* 286:35334-35338.

148. He YF, *et al.* (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* 333:1303-1307.

149. Cortazar D, *et al.* (2011) Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* 470:419-423.

150. Li YQ, Zhou PZ, Zheng XD, Walsh CP, & Xu GL (2007) Association of Dnmt3a and thymine DNA glycosylase links DNA methylation with base-excision repair. *Nucleic Acids Res* 35:390-400.

151. Tini M, *et al.* (2002) Association of CBP/p300 acetylase and thymine DNA glycosylase links DNA repair and transcription. *Mol Cell* 9:265-277.

152. Um S, *et al.* (1998) Retinoic acid receptors interact physically and functionally with the T:G mismatch-specific thymine-DNA glycosylase. *J Biol Chem* 273:20728-20736.

153. Hashimoto H (2014) Structural and mutation studies of two DNA demethylation related glycosylases: MBD4 and TDG. *Biophysics (Nagoya-shi)* 10:63-68.

154. Hashimoto H, Zhang X, & Cheng X (2012) Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: structural basis and implications for active DNA demethylation. *Nucleic Acids Res* 40:8276-8284.

155. Banavali NK & MacKerell AD, Jr. (2002) Free energy and structural pathways of base flipping in a DNA GCGC containing sequence. *J Mol Biol* 319:141-160.

156. Hitomi K, Iwai S, & Tainer JA (2007) The intricate structural chemistry of base excision repair machinery: implications for DNA damage recognition, removal, and repair. *DNA repair* 6:410-428.

157. Priyakumar UD & MacKerell AD, Jr. (2006) Computational approaches for investigating base flipping in oligonucleotides. *Chem Rev* 106:489-505.

158. Jacobs AL & Schar P (2012) DNA glycosylases: in DNA repair and beyond. *Chromosoma* 121:1-20.

159. Cao C, Jiang YL, Krosky DJ, & Stivers JT (2006) The catalytic power of uracil DNA glycosylase in the opening of thymine base pairs. *J Am Chem Soc* 128:13034-13035.

160. Cao C, Jiang YL, Stivers JT, & Song F (2004) Dynamic opening of DNA during the enzymatic search for a damaged base. *Nat Struct Mol Biol* 11:1230-1236.

161. Bruner SD, Norman DP, & Verdine GL (2000) Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature* 403:859-866.

162. Fromme JC & Verdine GL (2003) DNA lesion recognition by the bacterial repair enzyme MutM. *J Biol Chem* 278:51543-51548.

163. Qi Y, *et al.* (2009) Encounter and extrusion of an intrahelical lesion by a DNA repair enzyme. *Nature* 462:762-766.

164. Parker JB, *et al.* (2007) Enzymatic capture of an extrahelical thymine in the search for uracil in DNA. *Nature* 449:433-437.

165. Knips A & Zacharias M (2017) Both DNA global deformation and repair enzyme contacts mediate flipping of thymine dimer damage. *Sci Rep* 7:41324.

166. Maiti A, Morgan MT, Pozharski E, & Drohat AC (2008) Crystal structure of human thymine DNA glycosylase bound to DNA elucidates sequence-specific mismatch recognition. *Proc Natl Acad Sci U S A* 105:8890-8895.

167. Coey CT, *et al.* (2016) Structural basis of damage recognition by thymine DNA glycosylase: Key roles for N-terminal residues. *Nucleic Acids Res* 44:10248-10258.

168. Hamelberg D, de Oliveira CA, & McCammon JA (2007) Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J Chem Phys* 127:155102.

169. Perez-Hernandez G, Paul F, Giorgino T, De Fabritiis G, & Noe F (2013) Identification of slow molecular order parameters for Markov model construction. *J Chem Phys* 139:015102.

170. Metzner P, Schutte C, & Vanden-Eijnden E (2006) Illustration of transition path theory on a collection of simple examples. *J Chem Phys* 125:084110.

171. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, & Zakrzewska K (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* 37:5917-5929.

172. Hunter WN, *et al.* (1987) The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution. *J Biol Chem* 262:9962-9970.

173. Tsutakawa SE, Jingami H, & Morikawa K (1999) Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex. *Cell* 99:615-623.

174. Lamers MH, *et al.* (2000) The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch. *Nature* 407:711-717.

175. Obmolova G, Ban C, Hsieh P, & Yang W (2000) Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature* 407:703-710.

176. Zhu B, *et al.* (2000) 5-Methylcytosine DNA glycosylase activity is also present in the human MBD4 (G/T mismatch glycosylase) and in a related avian sequence. *Nucleic Acids Res* 28:4157-4165.

177. Qi Y, Spong MC, Nam K, Karplus M, & Verdine GL (2010) Entrapment and structure of an extrahelical guanine attempting to enter the active site of a bacterial DNA glycosylase, MutM. *J Biol Chem* 285:1468-1478.

178. Chung SJ & Verdine GL (2004) Structures of end products resulting from lesion processing by a DNA glycosylase/lyase. *Chem Biol* 11:1643-1649.

179. Nelson SR, Dunn AR, Kathe SD, Warshaw DM, & Wallace SS (2014) Two glycosylase families diffusively scan DNA using a wedge residue to probe for and identify oxidatively damaged bases. *Proc Natl Acad Sci U S A* 111:E2091-2099.

180. Dunn AR, Kad NM, Nelson SR, Warshaw DM, & Wallace SS (2011) Single Qdot-labeled glycosylase molecules use a wedge amino acid to probe for lesions while scanning along DNA. *Nucleic Acids Res* 39:7487-7498.

181. Vanden-Eijnden E & Venturoli M (2009) Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J Chem Phys* 130:194103.

182. Majek P & Elber R (2010) Milestoning without a Reaction Coordinate. *J Chem Theory Comput* 6:1805-1817.

183. Mathews DH & Case DA (2006) Nudged elastic band calculation of minimal energy paths for the conformational change of a GG non-canonical pair. *J Mol Biol* 357:1683-1693.

184. Kossmann B & Ivanov I (2014) Alkylpurine glycosylase D employs DNA sculpting as a strategy to extrude and excise damaged bases. *PLoS Comput Biol* 10:e1003704.

185. Slupphaug G, *et al.* (1996) A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA. *Nature* 384:87-92.

186. Parikh SS, *et al.* (2000) Uracil-DNA glycosylase-DNA substrate and product structures: conformational strain promotes catalytic efficiency by coupled stereoelectronic effects. *Proc Natl Acad Sci U S A* 97:5083-5088.

187. Parikh SS, *et al.* (1998) Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA. *EMBO J* 17:5214-5226.

188. Mol CD, *et al.* (1995) Crystal structure of human uracil-DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA. *Cell* 82:701-708.

189. Maier JA, *et al.* (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 11:3696-3713.

190. Pettersen EF, *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605-1612.

191. Moore, M.J. & Proudfoot, N.J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* 136, 688-700 (2009).

192. Proudfoot, N.J., Furger, A. & Dye, M.J. Integrating rnRNA processing with transcription. *Cell* 108, 501-512 (2002).

193. Bentley, D.L. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* 15, 163-175 (2014).

194. Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* 21, 327-35 (1996).

195. Zurita, M. & Cruz-Becerra, G. TFIIH: New discoveries regarding its mechanisms and impact on cancer treatment. *J Cancer* 7, 2258-2265 (2016).

196. Hishikawa, A., Hayashi, K. & Itoh, H. Transcription factors as therapeutic targets in chronic kidney disease. *Molecules* 23, 1123-1136 (2018).

197. Villard, J. Transcription regulation and human diseases. *Swiss Med Wkly* 134, 571-579 (2004).

198. Chen, X.F., Zhang, Y.W., Xu, H.X. & Bu, G.J. Transcriptional regulation and its misregulation in Alzheimer's disease. *Molecular Brain* 6, 44-53(2013).

199. Lee, T.I. & Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237-1251 (2013).

200. Goodrich, J.A., Cutler, G. & Tjian, R. Contacts in context: Promoter specificity and macromolecular interactions in transcription. *Cell* 84, 825-830 (1996).

201. Boeger, H. et al. Structural basis of eukaryotic gene transcription. *Febs Lett* 579, 899-903 (2005).

202. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P.A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* 56, 549-61 (1989).

203. Liu, X., Bushnell, D.A., Silva, D.A., Huang, X.H. & Kornberg, R.D. Initiation complex structure and promoter proofreading. *Science* 333, 633-637 (2011).

204. Cheung, A.C.M., Sainsbury, S. & Cramer, P. Structural basis of initial RNA polymerase II transcription. *Embo J* **30**, 4755-4763 (2011).

205. Sim, R.J., Belotserkovskaya, R. & Reinberg, D. Elongation by RNA polymerase II: the short and long of it. *Genes Dev* 18, 2437-2468 (2004).

206. He, Y. et al. Near-atomic resolution visualization of human transcription promoter opening. *Nature* 533, 359-365(2016).

207. Schilbach, S. et al. Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature* 551, 204-209 (2017).

208. He, Y., Fang, J., Taatjes, D.J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* 495, 481-486 (2013).

209. Greber, B.J. et al. The cryo-electron microscopy structure of human transcription factor IIH. *Nature* 549, 414-417 (2017).

210. Dubaele, S. et al. Basal transcription defect discriminates between xeroderma pigmentosum and trichothiodystrophy in XPD patients. *Mol Cell* 11, 1635-1646 (2003).

211. Coin, F. & Egly, J.M. Ten years of TFIIH. *Cold Spring Harb Symp Quant Biol* 63, 105-110 (1998).

212. Lehmann, A.R. The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases. *Genes Dev* 15, 15-23 (2001).

213. Berneburg, M. & Lehmann, A.R. Xeroderma pigmentosum and related disorders: defects in DNA repair and transcription. *Adv Genet* 43, 71-102 (2001).

214. Fassihi, H. et al. Deep phenotyping of 89 xeroderma pigmentosum patients reveals unexpected heterogeneity dependent on the precise molecular defect. *Proc Natl Acad Sci U S A* 113, E1236-E1245 (2016).

215. Boyle, J. et al. Persistence of repair proteins at unrepaired DNA damage distinguishes diseases with ERCC2 (XPD) mutations: cancer-prone xeroderma pigmentosum vs. non-cancer-prone trichothiodystrophy. *Hum Mutat* 29, 1194-208 (2008).

216. Rimel, J.K. & Taatjes, D.J. The essential and multifunctional TFIIH complex. *Protein Sci* 27, 1018-1037 (2018).

217. Compe, E. & Egly, J.M. Nucleotide excision repair and transcriptional regulation: TFIIH and Beyond. *Annu Rev Biochem,* 85, 265-290 (2016).

218. Singh, A., Compe, E., Le May, N. & Egly, J.M. TFIIH subunit alterations causing Xeroderma Pigmentosum and Trichothiodystrophy specifically disturb several steps during transcription. *Am J Hum Genet* 96, 194-207 (2015).

219. Compe, E. & Egly, J.M. TFIIH: when transcription met DNA repair. *Nat Rev Mol Cell Biol* 13, 343-354 (2012).

220. Grunberg, S. & Hahn, S. Structural insights into transcription initiation by RNA polymerase II. *Trends Biochem Sci* 38, 603-611 (2013).

221. Grunberg, S., Warfield, L. & Hahn, S. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nat Struct Mol Biol* 19, 788-96 (2012).

222. Fishburn, J., Tomko, E., Galburt, E. & Hahn, S. Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation. *Proc Natl Acad Sci U S A* 112, 3961-3966 (2015).

223. Singharoy, A. et al. Molecular dynamics-based model refinement and validation for sub-5 angstrom cryo-electron microscopy maps. *Elife* 5, e16105 (2016).

224. Luo, J. et al. Architecture of the human and yeast general transcription and DNA repair factor TFIIH. *Mol Cell* 59, 794-806 (2015).

225. Zhu, Q.Z., Wani, G., Sharma, N. & Wani, A. Lack of CAK complex accumulation at DNA damage sites in XP-B and XP-B/CS fibroblasts reveals differential regulation of

CAK anchoring to core TFIIH by XPB and XPD helicases during nucleotide excision repair. *DNA Repair* 11, 942-950 (2012).

226. Drapkin, R., LeRoy, G., Cho, H., Akoulitchev, S. & Reinberg, D. Human cyclin-dependent kinase-activating kinase exists in three distinct complexes. *Proc Natl Acad Sci USA* 93, 6488-6493 (1996).

227. Fuss, J.O. & Tainer, J.A. XPB and XPD helicases in TFIIH orchestrate DNA duplex opening and damage verification to coordinate repair with transcription and cell cycle via CAK kinase. *DNA Repair* 10, 697-713 (2011).

228. Coin, F., Oksenych, V. & Egly, J.M. Distinct roles for the XPB/p52 and XPD/p44 subcomplexes of TFIIH in damaged DNA opening during nucleotide excision repair. *Mol Cell* 26, 245-256 (2007).

229. Fan, L. et al. XPD helicase structures and activities: Insights into the cancer and aging phenotypes from XPD mutations. *Cell* 133, 789-800 (2008).

230. Fan, L. & DuPrez, K.T. XPB: An unconventional SF2 DNA helicase. *Prog Biophys Mol Biol* 117, 174-181 (2015).

231. Fan, L. et al. Conserved XPB core structure and motifs for DNA unwinding: Implications for pathway selection of transcription or excision repair. *Mol Cell* 22, 27-37 (2006).

232. Abdulrahman, W. et al. ARCH domain of XPD, an anchoring platform for CAK that conditions TFIIH DNA repair and transcription activities. *Proc Natl Acad Sci USA* 110, E633-E642 (2013).

233. Obmolova, G., Ban, C., Hsieh, P. & Yang, W. Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature* 407, 703-710 (2000).

234. Mason, A.C. et al. A structure-specific nucleic acid-binding domain conserved among DNA repair proteins. *Proc Natl Acad Sci USA* 111, 7618-7623 (2014).

235. Cheng, K. & Wigley, D.B. DNA translocation mechanism of an XPD family helicase. *Elife* 7**,** e42400 (2018).

236. Okuda, M. et al. Structural insight into the TFIIE-TFIIH interaction: TFIIE and p53 share the binding region on TFIIH. *Embo J* 27, 1161-1171 (2008).

237. Ohkuma, Y., Hashimoto, S., Wang, C.K., Horikoshi, M. & Roeder, R.G. Analysis of the role of TFIIE in basal transcription and TFIIH-mediated carboxy-terminal domain phosphorylation through structure-function studies of TFIIE-alpha. *Mol Cell Biol* 15, 4856-4866 (1995).

238. David, C.C. & Jacobs, D.J. Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol Biol* 1084, 193-226 (2014).

239. Parvin, J.D., Shykind, B.M., Meyers, R.E., Kim, J.S. & Sharp, P.A. Multiple sets of basal factors initiate transcription by RNA-Polymerase-II. *J Biol Chem* 269, 18414-18421 (1994).

240. Parvin, J.D. & Sharp, P.A. DNA Topology and a minimal set of basal factors for transcription by RNA Polymerase-II. *Cell* 73, 533-540 (1993).

241. Ueda, T., Compe, E., Catez, P., Kraemer, K.H. & Egly, J.M. Both XPD alleles contribute to the phenotype of compound heterozygote xeroderma pigmentosum patients. *J Exp Med* 206, 3031-3046 (2009).

242. DiGiovanna, J.J. & Kraemer, K.H. Shining a light on xeroderma pigmentosum. *J Invest Dermatol* 132, 785-796 (2012).

243. Kuschal, C. et al. GTF2E2 mutations destabilize the general transcription factor complex TFIIE in individuals with DNA repair-proficient Trichothiodystrophy. *Am J Hum Genet* 98, 627-42 (2016).

244. Kurowski, M.A. & Bujnicki, J.M. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res* 31, 3305-3307 (2003).

245. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60, 2126-2132 (2004).

246. Sali, A. & Blundell, T.L. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* 234, 779-815 (1993).

247. Radu, L. et al. The intricate network between the p34 and p44 subunits is central to the activity of the transcription/DNA repair factor TFIIH. *Nucleic Acids Res* 45, 10872-10883 (2017).

248. Gervais, V. et al. Solution structure of the N-terminal domain of the human TFIIH MAT1 subunit - New insights into the RING finger family. *J Biol Chem* 276, 7457-7464 (2001).

249. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66, 213-221 (2010).

250. Afonine, P.V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr D Struct Biol* 74, 531-544 (2018).

251. Kale, L. et al. NAMD2: Greater scalability for parallel molecular dynamics. *J Comput Phys* 151, 283-312 (1999).

252. Roe, D.R. & Cheatham, T.E., 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* 9, 3084-95 (2013).

253. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* 14, 33-8, 27-8 (1996).

254. Glykos, N.M. Software news and updates. Carma: a molecular dynamics analysis program. *J Comput Chem* 27, 1765-8 (2006).