



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

Ciclo 37

Settore Concorsuale: 03/D1 - CHIMICA E TECNOLOGIE FARMACEUTICHE,
TOSSICOLOGICHE E NUTRACEUTICO - ALIMENTARI

Settore Scientifico Disciplinare: CHIM/08 - CHIMICA FARMACEUTICA

UNRAVELING DNA TRANSLOCATION MECHANISM THROUGH DATA-DRIVEN
ATOMISTIC SIMULATIONS

Presentata da: Alessia Visigalli

Coordinatore Dottorato

Daniele Bonacorsi

Supervisore

Marco De Vivo

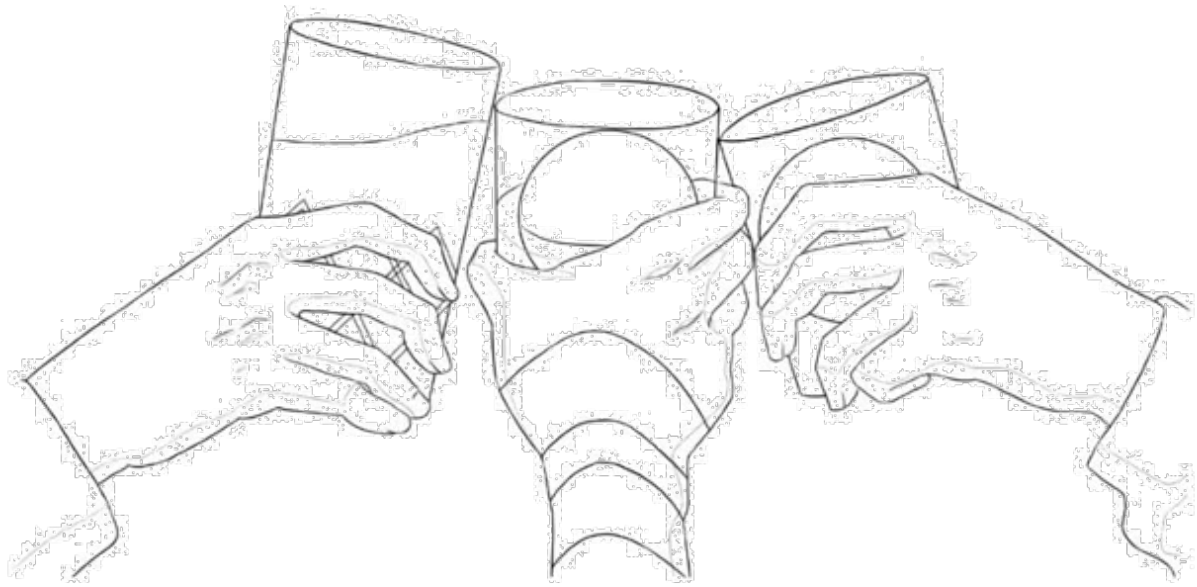
Co-supervisore

Andrea Cavalli

Esame finale anno 2026

*“Even if I were to live a thousand years, I would still want to
learn more. . .
see more. . .
understand more.”*

– Meruem



To the stray dogs.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to everyone who has supported me throughout my PhD journey. This accomplishment would not have been possible without your encouragement, guidance, and unwavering support.

I am deeply grateful to my supervisor, Dr. Marco De Vivo, for giving me the opportunity to work in his group and for guiding my growth as a scientist. I would also like to thank Dr. Pietro Vidossich for his invaluable support and the many scientific discussions that helped me improve my research. Thanks to all the members of the De Vivo group, past and present, for creating such a friendly and stimulating environment. A special thanks to Alice, Giorgia, and Stefano for their constant help and support throughout my PhD.

I am sincerely thankful to Prof. Dr. Paolo Carloni for hosting me in his group at the Institute of Computational Biomedicine in Jülich. My deepest appreciation also goes to Prof. Dr. Giulia Rossetti for her scientific guidance and personal support during my stay in Jülich. Thanks to all the members of the Carloni group, past and present, for the welcoming and inspiring atmosphere.

To my longtime friends, thank you for always being there for me, even when we are miles apart. A special mention goes to Martina and Alice, for your endless support and friendship.

To my parents and my brother, thank you for your constant love, patience, and encouragement throughout my studies and my life. And to Alessandro, thank you for being love, support, colleague, and friend all in one.

To my colleagues at the University of Bologna. Despite the distance and the pandemic, we share the same journey, joy and pain.

Finally, I would like to thank my former colleagues at the University of Parma, where I completed my Bachelor's and Master's studies, Lorenzo, Matteo, Stefano, and Vincenzo, for the great time we shared. A special thanks to Prof. Dr. Stefania Abbruzzetti for her guidance and for inspiring me to pursue this PhD path.

Thank you all for being part of this incredible journey.

Abstract

DNA replication is an essential and tightly regulated process that guarantees the faithful transmission of genetic information in all living organisms. This fundamental mechanism relies on the coordinated action of numerous proteins that ensure both the accuracy and efficiency of DNA synthesis. Among them, DNA polymerases (Pols) play a central role by catalyzing the addition of nucleotides to the growing DNA strand, using the complementary template as a guide.

Replicative polymerases are responsible for the bulk of genome duplication and are characterized by high fidelity, proofreading activity, and remarkable processivity, which together ensure the correct copying of the entire genome. However, mutations in these enzymes or alterations in their regulation can compromise genomic stability, leading to replication errors, mutagenesis, and uncontrolled cell proliferation. In contrast, specialized translesion synthesis (TLS) Pols are capable of bypassing DNA lesions that would otherwise stall the replication machinery. These lesion bypass polymerases, often belonging to the Y-family, possess more flexible active sites that allow them to accommodate damaged or distorted DNA templates. For instance, ultraviolet (UV) radiation induces characteristic DNA lesions such as cyclobutane pyrimidine dimers (CPD), which are substrates for TLS polymerases. While these enzymes play a crucial protective role by preventing replication fork collapse, their inherently low fidelity can also promote mutagenesis when misregulated. Owing to their pivotal role in genome maintenance, both replicative and lesion bypass polymerases have emerged as attractive targets for small-molecule modulation in cancer therapy.

DNA polymerases are metalloenzymes that catalyze the faithful copying of single-stranded DNA through a well-known two-metal-ion mechanism. Following a stepwise catalytic cycle, Pols incorporate an incoming deoxynucleotide triphosphate (dNTP) at the 3' end of the growing primer strand. After each catalytic event, the DNA substrate translocates by one base pair, repositioning the primer terminus within the active site for the next binding event. Although the chemical step of catalysis has been extensively characterized using both experimental and computational approaches, the molecular details of the translocation process remain poorly understood at the atomistic level.

Recent computational studies have elucidated the translocation mechanism in replicative Pols, revealing that a coordinated open-to-close transition of a structural helix located above the DNA duplex facilitates DNA movement away from the active site. Intriguingly, lesion bypass Pols lack this structural element and display no major conformational rearrangements throughout the entire catalytic cycle. This raises fundamental questions about how these enzymes achieve DNA translocation and maintain catalytic efficiency despite their distinct structural organization and functional specialization.

Aim of the Thesis

The aim of this PhD thesis is to elucidate the molecular mechanism that governs the translocation step in lesion bypass polymerases, employing molecular dynamics (MD) simulations and enhanced sampling techniques. To this end, we focused on DNA polymerase η (Pol η) as a representative model of Y-family polymerases involved in translesion synthesis (TLS). Pol η plays a crucial role in maintaining genome integrity by enabling DNA synthesis across damaged templates, and its dysfunction has been linked to cancer development and therapy resistance.

The first chapter introduces the biological background of this work, emphasizing the role of DNA polymerases in replication and repair, with a particular focus on the structural and functional properties of Pol η , its catalytic cycle, and the current understanding of the translocation process. The second chapter presents the computational methodologies employed in this study, highlighting both classical molecular dynamics and enhanced sampling simulations, as well as the use of machine learning-based collective variables.

In the third chapter, we investigate the translocation mechanism in the wild-type Pol η system, elucidating the crucial role of electrostatic interactions between the protein and DNA. The fourth chapter explores the translocation defects induced by the mutation of positively charged residues at the protein-DNA interface, combining computational insights with experimental validation. The fifth chapter examines the structural and functional consequences of a pathogenic variant of Pol η associated with a genetic disorder. Finally, the sixth chapter summarizes the main findings of this thesis and outlines potential directions for future research.

Table of Contents

Acknowledgements	ii
Abstract	iv
Aim of the Thesis	v
1 Introduction	1
1.1 The Central Role of Nucleic Acids in Life	1
1.1.1 Structure and function of DNA and RNA	1
1.1.2 DNA replication: a fundamental process for life	2
1.2 Polymerases: Structure, Function, and Biological Significance	3
1.2.1 Overview of DNA and RNA polymerases	3
1.2.2 Pharmacological and biotechnological relevance	4
1.3 Structural and Functional Features of DNA Polymerases	5
1.3.1 General architecture of DNA polymerases	5
1.3.2 Fidelity, processivity, and polymerase classification	7
1.3.3 Y-family: Translesion Synthesis Polymerases	12
1.4 Pols Catalytic Cycle and Translocation Mechanism	15
1.4.1 DNA Translocation Mechanism	18
1.5 Specialized DNA Pols and DNA Damage Bypass	20
1.5.1 Types of DNA Lesions	20
1.5.2 Role of Y-family DNA Polymerases in DNA Damage Tolerance	22
2 Materials and Methods	25
2.1 Molecular Dynamics Simulations	25
2.1.1 Classical Equation of Motion	25
2.1.2 Potential energy and Force fields	26
2.1.3 Hamiltonian and Lagrangian formalism	29
2.1.4 Statistical ensembles	30
2.1.5 Ergodicity in MD simulations	32
2.1.6 Numerical Integration of the equations of motion	32
2.1.7 Temperature and pressure control	33
2.1.8 Explicit solvent and PBC	34
2.2 Sampling the Rare Events	35

2.2.1	Enhanced Sampling Methods	35
2.2.2	Towards collective variable space	35
2.2.3	On-the-fly Probability Enhanced Sampling (OPES)	38
2.3	Machine learning based collective variables	39
2.3.1	Different classes of Machine Learning methods	40
2.3.2	Data-driven approaches to CV design	41
3	DNA Translocation in wild-type Polη	45
3.1	Introduction	46
3.2	Identification of Key Residues Promoting DNA Translocation	47
3.2.1	Primer•Pol η Interactions	47
3.2.2	Template•Pol η Interactions	48
3.3	Characterization of the Asynchronous Stepwise Mechanism.	49
3.3.1	Implementing an Optimized Collective Variable for DNA Translocation.	51
3.4	Discussion	52
3.5	Conclusions	54
4	Arginine Mutations: Functional Role in Translocation	56
4.1	Introduction	57
4.2	Experimental Validation of Arginine Mutant Activity and Processivity	57
4.3	Computational insights into Arginine Mutations in DNA translocation	59
4.3.1	Arg377Ala: Structural Compensation and DNA Misalignment	59
4.3.2	Arg382Ala and Arg383Ala: Differential Roles in DNA Stabilization and Translocation	60
4.3.3	Arg382Ala/Arg383Ala double mutant promotes DNA unbinding	61
4.4	Discussion	62
5	Impact of Pathological Mutation on Polη Translocation	64
5.1	Introduction	65
5.2	Ala117Pro and Thr122Pro mutants disrupt the catalytic site	66
5.3	Gly263Val, Arg93Pro, and Arg111His affect translocation and DNA binding	67
5.4	Arg361Ser impairs the stability of the little-finger subdomain	68
5.5	Discussion	69
6	Conclusions and Perspectives	71
A	Supplementary Material for Chapter 3	73
A.1	Material and Methods	74
A.1.1	Structural Models Preparation.	74
A.1.2	Classical MD Simulations.	74
A.1.3	Enhanced Sampling Simulations.	75
B	Supplementary Material for Chapter 4	86
B.1	Material and Methods	87

B.1.1	Structural Models Preparation.	87
B.1.2	Molecular Dynamics Simulations	87
B.1.3	Clustering Analysis.	87
C	Supplementary Material for Chapter 5	95
C.1	Structural Models Preparation.	96
C.2	Simulation Protocol.	96
C.3	Data and Software Availability.	96
	Bibliography	101

List of Figures

1.1	Structural comparison between DNA and RNA.	1
1.2	Schematic representation of the DNA replication process in eukaryotes.	2
1.3	General architecture of DNA polymerases.	5
1.4	Structural comparison of two DNA polymerases with additional subdomains.	6
1.5	Crystal structures of the Y-family members	12
1.6	Spotlight on the structural gap between the polymerase core and little finger (LF) subdomain in Y-family polymerases.	14
1.7	Structures of Dpo4 in apo and binary complex	15
1.8	Stepwise catalytic cycle of DNA polymerases.	16
1.9	Schematic representation of the catalytic mechanism the two-metal-ion mechanism in DNA and RNA polymerases.	17
1.10	Conformational changes of the fingers subdomain replicative Pols	19
1.11	DNA lesions	20
1.12	Crystal structure of human Pol η in the presence of XP-V related mutations	23
2.1	Schematic representation of enhanced sampling methods	37
2.2	Overview of different neural network architectures	40
2.3	Overview of different data-driven approaches for collective variable (CV) design in molecular dynamics simulations.	41
2.4	Example of multitask workflow for collective variable (CV) design using semi-supervised learning.	43
3.1	Schematic representation of the Pol η catalytic cycle and Crystal structure of the DNA polymerase η	46
3.2	Overview of positive residues at the Pol η •DNA interface	48
3.3	Schematic representation of the Pol η •DNA interactions	49

3.4	Description and classification of the two intermediate states	50
3.5	Free energy surface describing the translocation process in wild-type Pol η	51
3.6	Schematic representation of the asynchronous DNA translocation mechanism in Pol η	53
3.7	Structural comparison of positively charged residues interacting with DNA in Y-family polymerases	54
4.1	Preliminary experimental results of Pol η mutated systems	58
4.2	Arg377 in wild-type Pol η	59
5.1	Structural overview of Ala117 and Thr122 in wild-type Pol η	66
5.2	Structural overview of Gly263, Arg93, and Arg111 in wild-type Pol η	67
5.3	Structural overview of Arg361 in wild-type Pol η	69
A.1	Pre-translocation state heavy atoms RMSD	77
A.2	Post-translocation state heavy atoms RMSD	78
A.3	Pre-translocation state DNA residue-DNA distances	79
A.4	Pre-translocation state DNA RMSD	80
A.5	Evolution of the On-the-fly Probability Enhanced Sampling (OPES) 1D collective variable (CV) over time.	81
A.6	Root-Mean-Square-Deviation (RMSD) of the template strand (blue) and the primer strand (red) over time during the translocation.	82
A.7	Classical MD simulations of the two intermediate states.	83
A.8	Aromatic residues at the finger•little-finger interface in Y-family Pols.	84
A.9	Selection of the descriptors for the machine-learning CVs	85
B.1	Switching Function	88
B.2	Root-Mean-Square-Deviation (RMSD) of the protein backbone in Arg377Ala mutant system.	89
B.3	Density function of descriptors involving primer strand bases in Arg377Ala mutant system in pre-translocation state.	90
B.4	Root-Mean-Square-Deviation (RMSD) of the protein backbone in Arg382Ala mutant system.	91
B.5	Root-Mean-Square-Deviation (RMSD) of the protein backbone in Arg383Ala mutant system.	92
C.1	Rearrangement of active site hydrogen bonds in Ala117Pro	97
C.2	Rearrangement of active site hydrogen bonds in Thr122Pro	98
C.3	Changing in Ramachandran plot of Arg93Pro	99
C.4	Steric gate network in wild-type Pol η	100

Chapter I – Introduction

1.1 The Central Role of Nucleic Acids in Life

1.1.1 Structure and function of DNA and RNA

DNA and RNA are long, linear polymers classified as nucleic acids, which serve as the primary carriers of genetic information in all known living organisms. These macromolecules are composed of repeating units called nucleotides, each consisting of a nitrogenous base, a five-carbon sugar (deoxyribose in DNA, ribose in RNA), and one or more phosphate groups. DNA typically adopts a double-helical structure, in which two antiparallel strands are held together by hydrogen bonds between specific base pairs: adenine (A) with thymine (T), and cytosine (C) with guanine (G). These canonical Watson-Crick (WC) base pairs are stabilized by hydrogen bonds ($\sim 1\text{-}5 \text{ kcal mol}^{-1}$) and further reinforced by base stacking interactions, primarily van der Waals forces ($\sim 0.5\text{-}1 \text{ kcal mol}^{-1}$ per pair), which contribute significantly to the overall stability of the double-helical structure.¹ Although these non-covalent interactions are individually weak, their cumulative effect over hundreds or thousands of base pairs results in a highly stable and compact structure, well-suited to long-term information storage.²

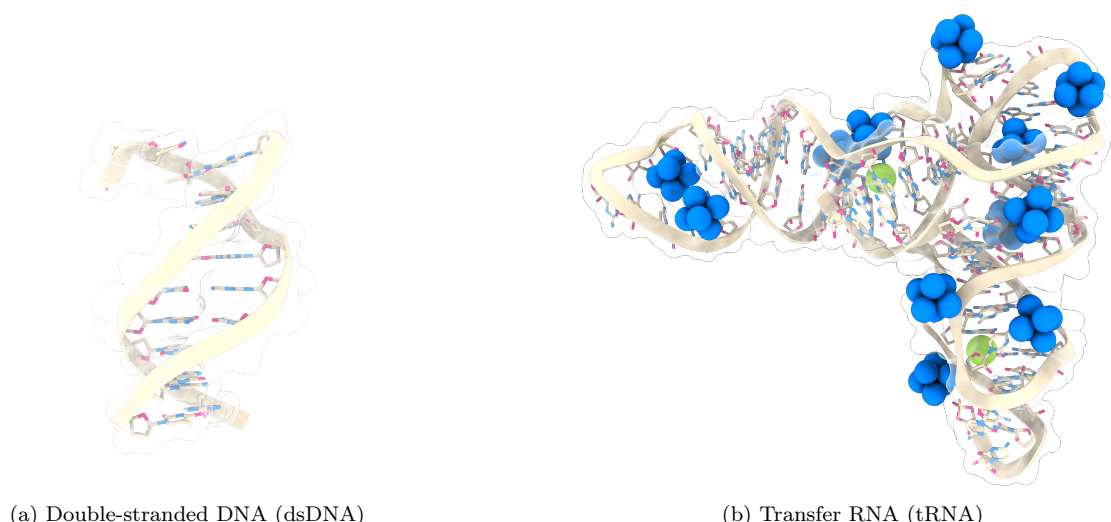


Figure 1.1: **Structural comparison between DNA and RNA.** (a) Double-stranded DNA helix (PDB ID: 4ECX), stabilized by Watson-Crick base pairing and base stacking interactions. (b) Transfer RNA (tRNA) secondary structure (PDB ID: 4P5J), illustrating the versatility of RNA folding driven by intramolecular base pairing. The figure was created using Chimera X molecular graphics system.³

Conversely, RNA is usually single-stranded and often folds into complex three-dimensional structures driven by intramolecular base pairing and tertiary interactions. While RNA can adopt helical regions similar to DNA, it exhibits greater structural versatility and is involved in a broad range of cellular functions beyond genetic storage. These include roles as messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and catalytic RNA species (ribozymes), as well as regulatory non-coding RNAs.² One key chemical difference between RNA and DNA is the presence of a hydroxyl group at the 2'

position of the ribose sugar in RNA, which renders RNA more reactive and less chemically stable than DNA, particularly under alkaline conditions. Together, DNA and RNA play complementary roles in the central dogma of molecular biology: DNA stores genetic information, which is transcribed into RNA and ultimately translated into proteins. Their structure-function relationship, based on the specificity of base pairing and the ability to form stable secondary and tertiary structures, underlies the fidelity and regulation of genetic processes such as replication, transcription, and translation.²

1.1.2 DNA replication: a fundamental process for life

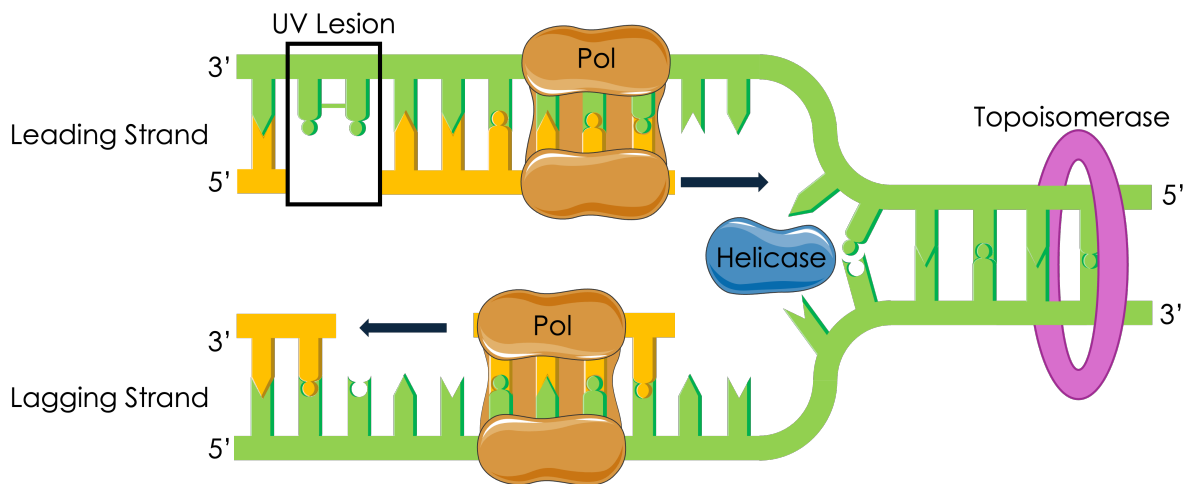


Figure 1.2: **Schematic representation of the DNA replication process in eukaryotes.**² At the replication fork, the topoisomerase enzyme (purple) relieves torsional stress generated ahead of the fork by introducing negative supercoils into the DNA, thus facilitating fork progression. The helicase enzyme (blue) unwinds the DNA double helix, creating single-stranded regions that serve as templates for replication. The parent strand is divided into two complementary template strands (green). Since the DNA replication proceeds from the 3' to the 5' terminal, the leading strand is synthesized continuously (up) to a primer strand (yellow) by the DNA polymerases (brown). Whereas the lagging strand is synthesized discontinuously (down) in short fragments (Okazaki fragments) by another DNA polymerase (brown).⁴ The navy-blue arrow indicates the direction of the DNA replication. A possible UV lesion is shown in the lagging strand (black rectangle), which is bypassed by translesion synthesis (TLS) polymerases (not indicated in this scheme).⁵ Adapted from Servier Medical Art (<https://smart.servier.com>), licensed under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

DNA replication is a crucial process that ensures the faithful transmission of genetic information from one generation to the next.⁶ It is the mechanism by which a cell duplicates its DNA prior to cell division, allowing each daughter cell to inherit a complete set of genetic instructions. The accuracy and efficiency of DNA replication are vital for maintaining genomic stability and preventing mutations that could lead to diseases such as cancer. The full replication process in a mammalian cell involves multiple steps and more than 20 enzymes, including DNA and RNA polymerases (Pols). This fascinating and intricate mechanism can be broadly divided into three main phases: initiation, elongation, and termination, each characterized by the coordinated action of various proteins.²

Initiation phase

In this phase, the DNA is unwound, and the replicative process begins. Originally, the DNA is packaged into nucleosomes with histone proteins. DNA replication starts at specific sites called origins of replication.² In yeast, these origins have defined sequences, while in humans, replication origins are likely recognized through specific nucleosome modifications rather than sequence. Initiator proteins bind to these origins and recruit other replication proteins, including helicases that unwind the DNA to form replication forks (see Figure 1.2). Two replication forks form at each origin, expanding in both directions to create a replication bubble.² Multiple origins on each chromosome allow replication to proceed simultaneously at many sites.

Elongation phase

This phase starts with the synthesis of ~ 10 nucleotides of the primer strand by an RNA polymerase called primase.² Both DNA strands serve as templates for the new DNA synthesis. Then, two DNA Pols shift along the two template strands, synthesizing two complementary primer filaments (see Figure 1.2). To do that, DNA Pols add a new nucleotide to the 3' end of the growing strand, proceeding in the $5' \rightarrow 3'$ direction (see yellow filaments and navy-blue arrows in Figure 1.2).⁷ Notably, the two template strands are antiparallel, which means that one primer strand appears to grow in the opposite direction ($3' \rightarrow 5'$). This primer is called the lagging strand, and it is synthesized in small pieces, known as Okazaki fragments.² By contrast, the other primer is synthesized continuously and is known as the leading strand.⁴ While the two single template strands are maintained open by the DNA helicases, other proteins like single-strand binding proteins and topoisomerases stabilize the open shape and prevent the supercoiling aggregation (see Figure 1.2).²

Termination phase

As replication proceeds, more and more nucleotides are added to synthesize the primer strand. Nevertheless, DNA Pols are not able to join the final phosphodiester bonds between adjacent fragments, leaving nicks in the sugar-phosphate backbone.² So then, enzymes like FEN1 and RNase H remove the RNA primer fragment, the DNA Pols fill the gaps, and finally the DNA ligase joins the pieces.⁸

1.2 Polymerases: Structure, Function, and Biological Significance

1.2.1 Overview of DNA and RNA polymerases

Nucleic acid polymerases are pivotal enzymes involved in DNA and RNA synthesis.⁹ These fascinating proteins are responsible for different biological mechanisms involving nucleic acid, such as DNA or RNA replication (DNA $\rightarrow$ DNA or RNA $\rightarrow$ RNA), transcription (DNA $\rightarrow$ RNA), reverse transcription (RNA $\rightarrow$ DNA), and DNA repair.⁷ Therefore, mutated versions of those systems are involved in several genetic diseases, including bacterial infections and cancer.^{10–13} Furthermore, viruses exploit polymerase enzymes for their own replication. For instance, DNA viruses often use host DNA polymerases (e.g., papillomaviruses), though larger viruses (e.g., herpesviruses) encode their own enzyme.² RNA-based viruses

encode RNA-dependent RNA polymerases (RdRp), since host cells lack enzymes capable of replicating RNA from RNA templates.¹⁴ Retroviruses like the Human Immunodeficiency Virus (HIV) use reverse transcriptase (RT), an RNA-dependent DNA polymerase, to convert their RNA genome into DNA, which is then integrated into the host genome.^{2,15} These viral polymerases are key targets for antiviral drugs (e.g., remdesivir targets RdRp in SARS-CoV2, while AZT inhibits HIV reverse transcriptase).^{16–19} In the past decades, multiple studies focused on the RNA polymerase of SARS coronavirus 2 (SARS-CoV-2) in order to find a possible treatment for coronavirus disease 19 (COVID-19).²⁰ For instance, viral vector vaccines against SARS-CoV-2 exploit the mechanism of human RNA polymerase to transcribe the virus genetic material.²¹

For these reasons, RNA and DNA Pols are promising pharmacological targets for small-molecules inhibitors.^{17,22} Polymerase inhibitors are clinically used as antibiotics (e.g., rifampicin) and antivirals (e.g., favipiravir and remdesivir), targeting the RdRp of bacteria and viruses, respectively.^{16,18} Notably, Pols are also implied in biotechnological applications, including the polymerase chain reaction (PCR) method discovered by Kary B. Mullis (Nobel Prize in Chemistry, 1993).²³ Reverse transcriptases are used in RT-PCR for gene expression analysis and in the development of mRNA vaccines, while phage RNA polymerases (like T7) are used in vitro transcription systems for producing synthetic RNA.^{21,24}

1.2.2 Pharmacological and biotechnological relevance

Studying these enzymes through computational methods is of vital importance, as it allows us to understand their mechanisms at the level of catalysis and translocation.^{7,25–28} Furthermore, such studies are essential for the development of targeted drugs and engineered enzymes.^{17,22} Polymerases are in fact involved in a wide range of therapeutic, biotechnological, and epigenetic applications.⁷

Drug design

Polymerases represent highly relevant pharmacological targets for all the domains of life. In the past, computational approaches, including molecular docking and free energy perturbation (FEP) have been successfully employed to identify effective inhibitors of some of these enzymes, such as the polymerase involved in HIV replication.¹⁵ More recently, efforts have focused on the development of advanced drugs selectively targeting non-replicative polymerases, those specialized in translesion synthesis, including Pol η , Pol κ , Rev1, and Pol ι .²² This growing interest is driven by the discovery of a correlation between the dysfunction of these enzymes and the onset of certain types of cancer.^{29,30} Moreover, the ability of some of those enzymes to bypass DNA lesions has been associated with reduced efficacy of specific chemotherapeutic agents, as they can accurately replicate DNA strands in the presence of drug-induced nucleotide adducts.^{31,32} As a result, there is a need for the development of selective compounds capable of inhibiting or enhancing the activity of specific polymerases, depending on the therapeutic context.¹⁷ Another pharmacological target of considerable recent interest is the RNA-dependent RNA polymerase of SARS-CoV-2, the virus responsible for COVID-19 epidemic. Numerous computational and experimental studies have been conducted to elucidate its mechanism of action and to facilitate the design of drugs aimed at inhibiting its activity.^{16,18,22} These studies have provided valuable insights into the enzyme structure and function, paving the way for the development of effective antiviral therapies.

Biotechnology and Epigenetics

One of the major applications of polymerases is the polymerase chain reaction (PCR), discovered in 1984 by Kary Mullis.²³ This fundamental technique allows the amplification of DNA sequences, providing valuable diagnostic information in medicine. For instance, PCR can reveal the presence of the HIV DNA even in patients who have not yet developed symptoms.² Also, this great amplification capacity can be highly informative in monitoring cancer chemotherapy. PCR tests can detect when cancerous cells have been destroyed, and the treatment can be stopped. Additionally, PCR can be utilized in forensics and legal medicine to establish biological parentage or prove guilt or innocence in various crimes.² PCR is also used in epigenetic studies, such as the analysis of DNA methylation, which plays a key role in gene regulation and is often altered in diseases like cancer.²⁴

1.3 Structural and Functional Features of DNA Polymerases

1.3.1 General architecture of DNA polymerases

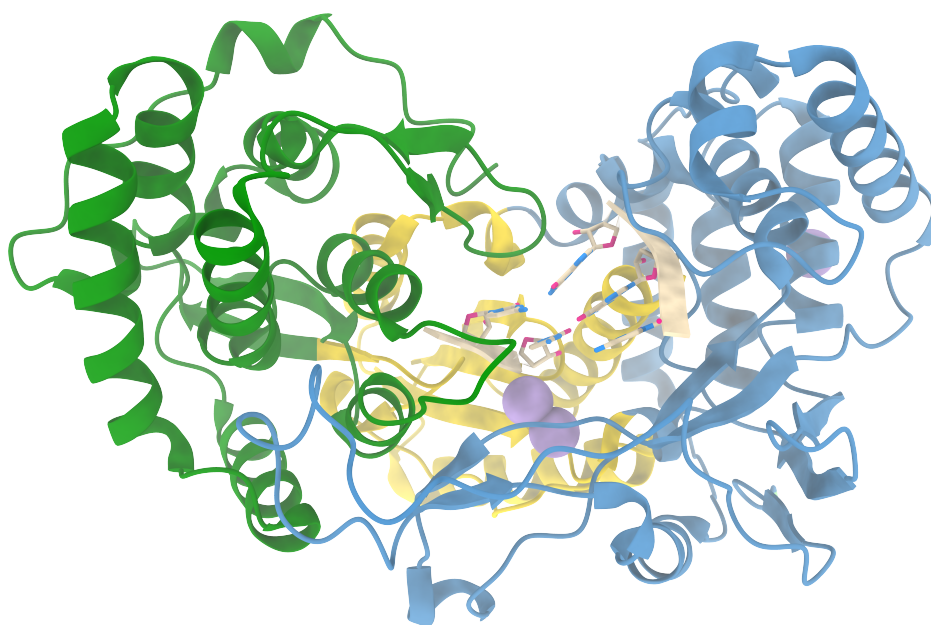


Figure 1.3: **General architecture of DNA polymerases.** The catalytic core resembles a right hand with three subdomains: palm (gold), fingers (lightblue), and thumb (green). The double-stranded DNA is shown in wheat, and the manganese ions present in the active site are in purple spheres. The palm contains the active site, while the fingers and thumb subdomains stabilize the DNA and facilitate catalysis.^{7,33,34} The figure shows the catalytic core of a representative and very simple RNA polymerase (Hepatitis C RNA polymerase, PDB ID: 4WTI).¹⁴ The catalytic core is highly conserved across different polymerase families, despite variations in sequence and additional accessory subdomains.^{7,25} The figure was created using Chimera X molecular graphics system.³

Despite the differences in sequence and functions, DNA and RNA polymerases share many structural features across all the families. Indeed, Pols are characterized by the common catalytic core that recalls a right hand, formed by three subdomains: "palm", "finger", and "thumb" (Figure 1.3).^{33,34} In most cases, these subdomains have the same function, but their topology and ternary structure differ across the families.⁷ The palm subdomain contains the active site, composed of three negatively charged residues (catalytic triad), which coordinate the two (or three) magnesium ions (Mg^{2+}) at the catalytic center.⁷ The finger subdomain lies upon the active site, with the highly flexible and conserved positive residue (Arg or Lys) stabilizing the catalytic transition state and promoting the exit of the leaving group pyrophosphate (PPi) after nucleotide addition.³⁵ Notably, the flexible helix present in the finger subdomain of the majority of DNA Pols is able to rotate before and after the catalysis, closing the active site and promoting the DNA translocation (see subsection 1.4.1 for more details).^{7,36,37} Finally, the thumb subdomain surrounds the double-strand DNA, anchoring it through positively charged residues at the Pol•DNA interface.

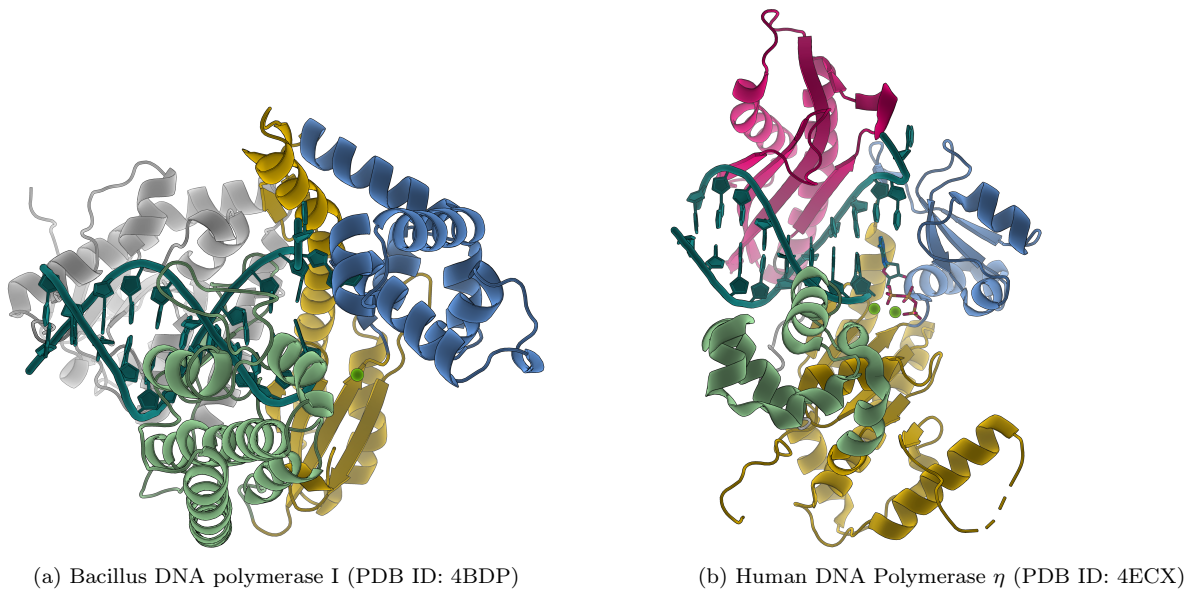


Figure 1.4: **Structural comparison of two DNA polymerases with additional subdomains.** (a) Bacillus DNA polymerase I (PDB ID: 4BDP), which contains a 3'-5' exonuclease subdomain (white) needed for proofreading activity.³⁸ (b) Human DNA Polymerase η (PDB ID: 4ECX), which includes a little-finger (LF) subdomain (magenta) needed by Y-family Pols to better bind the DNA substrate and enhance processivity and translocation.^{5,27,34,39} The catalytic core is shown in gold (palm), light blue (fingers), and light green (thumb). The DNA is depicted in teal. The figure was created using the PyMOL 3 molecular graphics system.⁴⁰

These three subdomains not only define the structural scaffold of the enzyme, but also participate in key conformational changes essential for catalytic efficiency and fidelity.²⁵ In replicative polymerases, the movement of the fingers subdomain upon correct nucleotide binding facilitates a closed conformation, which aligns the reactive groups for phosphoryl transfer and serves as a fidelity checkpoint.^{7,41} Conversely, in translesion synthesis (TLS) polymerases, the fingers subdomain is generally shorter and more flexible, reflecting a trade-off between catalytic efficiency and the ability to accommodate distorted or damaged DNA templates.⁵ Indeed, Y-family Pols compensate for the absence of a wider finger subdomain with the presence of the additional little-finger subdomain.⁵ The thumb subdomain, while primarily in-

volved in binding and positioning the DNA duplex, also contributes to enzyme processivity by stabilizing the Pol η •DNA complex throughout consecutive nucleotide additions.⁴² Notably, the palm, fingers, and thumb subdomains represent the most conserved structural elements within the catalytic core of all DNA and RNA polymerases.² However, many polymerases also contain additional subdomains or terminal extensions outside this core, which confer specialized functions such as proofreading exonuclease activity, processivity enhancement, or recognition of post-translational modifications.^{7,43} Examples include the 3'-5' exonuclease subdomain in high-fidelity polymerases, the PAD (polymerase-associated domain) or little-finger (LF) in Y-family enzymes, and ubiquitin-binding motifs (e.g., UBZ or UBM subdomains) in translesion polymerases.⁴³ These accessory portions of the protein expand the functional versatility of polymerases and are often tailored to the biological context in which each enzyme operates.

1.3.2 Fidelity, processivity, and polymerase classification

DNA and RNA polymerases are characterized by two key enzymatic properties: fidelity and processivity.⁷ Both of them are crucial for maintaining genome integrity and ensuring efficient DNA replication.⁷ Fidelity refers to the accuracy of nucleotide incorporation, while processivity refers to the continuity of DNA synthesis. Understanding and characterizing these properties is essential for elucidating the mechanisms underlying DNA replication and repair, as well as for developing targeted inhibitors or enhancers of polymerase activity.^{7,22} Both properties are influenced by several factors, including the enzyme structural features, catalytic mechanisms, and interactions with DNA and other cofactors.⁷ Both can vary significantly among different polymerases, reflecting their specialized roles in cellular processes such as replication, repair, and translesion synthesis (TLS).⁷ In this subsection, we will explore the definitions of fidelity and processivity, their significance in DNA replication, and how they relate to the classification of DNA and RNA polymerases.

Fidelity

Fidelity refers to the enzyme ability to accurately incorporate the correct nucleotide that is complementary to the template base during DNA synthesis. High-fidelity polymerases exhibit a low error rate (typically between 10^{-5} and 10^{-6}), often enhanced by 3'-5' exonuclease (proofreading) activity that removes misincorporated nucleotides. This ensures faithful genome duplication and minimizes mutagenesis.⁷

$$\text{Fidelity} = \frac{\left(\frac{k_{\text{pol}}}{K_d}\right)_{\text{correct}}}{\left(\frac{k_{\text{pol}}}{K_d}\right)_{\text{incorrect}}} \quad (1.1)$$

Therefore, Pols have different mechanisms by which can prevent the incorporation of an incorrect nucleotide.²⁸ Notably, the incorporation of an RNA nucleotide is prevented by the steric clash between the 2'-OH group on the ribose ring of the incoming nucleotide and the "steric gate" residue. This well-conserved amino acid (Glu for A-family Pols, Tyr or Phe for all the other families), is located in the finger subdomain, behind the insertion site.⁴⁴⁻⁴⁸ Indeed, several experimental studies highlight the reduced sugar discrimination of different Pols in the presence of a mutation involving the steric gate residue.⁴⁷ Despite this mechanism being well conserved among all the families, every enzyme shows a

different fidelity ratio in the presence of ribonucleotide or base mismatch.

Another intriguing selection mechanism is related to the presence of a positively charged residue (Arg or Lys) at the insertion site.³⁵ It was proposed that the presence of this highly conserved positive residue prevents the formation of a Hoogsteen base pair in Pol η , establishing electrostatic interactions with the incoming nucleotide, promoting its stability.^{33,35} Finally, it was proposed for yeast Pol η that the absence of the conserved Tyr52, located above the insertion site in the finger subdomain, can alter the fidelity of this enzyme.⁴⁹ This tyrosine could be used by the enzyme to maintain the incoming nucleotide in the correct position, since this enzyme lacks the mobile helix typical of high-fidelity polymerases, but also prevents the mismatch maintaining the correct geometry of the active site. Notably, the same mechanism involving the conserved Tyr was also found in other Pols.^{50–52} All together, these insights suggest that the main control for fidelity in Pols concerns the finger subdomains, which are able to stabilize the correct nucleotide in the active site and prevent the incorporation of an incorrect one.

Processivity

Processivity describes the number of nucleotides added by a polymerase in a single binding event, which means before dissociating from the DNA substrate.⁷ Highly processive enzymes can synthesize long stretches of DNA in a single binding event, which is essential for efficient genome replication (>100 nt/s).⁷ Low-processivity polymerases dissociate more frequently and are typically involved in DNA repair or lesion bypass, where short synthesis tracts are sufficient or required (<20 nt/s).⁵ Replicative polymerases (e.g., Pol δ , Pol ϵ in eukaryotes; PolIII in prokaryotes) exhibit both high fidelity and high processivity.⁷ By contrast, translesion synthesis (TLS) and repair polymerases, such as members of the Y-family, are characterized by lower fidelity and reduced processivity, enabling them to bypass damaged DNA or perform limited repair synthesis.⁵ This trade-off reflects their specialized roles in maintaining genome integrity under stress conditions.

$$\text{Processivity} = \frac{\# \text{ incorporated nucleotide}}{\# \text{ dissociation events}} \quad (1.2)$$

Despite the literature lacking experimental studies regarding processivity, this property is typically estimated as the probability of incorporating an additional nucleotide rather than dissociating from the template (see equation 1.3).⁵³

$$P_n = \frac{I_{n+1} + I_{n+2} + \dots}{I_n + I_{n+1} + I_{n+2} + \dots} \quad (1.3)$$

Notably, some DNA polymerases rely on accessory subdomains or cofactors, such as PCNA, to enhance their processivity.⁴³ For instance, recent studies investigating the processivity of Pol κ have demonstrated a marked reduction in this property when the enzyme is truncated, meaning it contains only the four canonical structural subdomains of Y-family polymerases (palm, fingers, thumb, and little finger) but lacks additional regulatory regions, such as the C-terminal extensions involved in PCNA interactions (e.g., the PIP box).⁵⁴

Polymerases	Processivity
yPol η ^{53,55}	~ 5
hPol η ⁵⁶	~ 8
hPol ι ⁵⁷	~ 1
hPol κ (Pols domain) ⁵⁴	~ 20
hPol κ (full-length) ⁵⁴	>25
Rev1 ⁵⁸	~ 1 or 2
Pol IV ⁵⁵	~ 15
Pol V ⁵⁹	~ 30

Table 1.1: **The processivity of each Y-family Pol** is reported in the table, along with the reference for the experimental data. Note that the values are approximate and can vary depending on the experimental conditions and specific assays used to measure processivity.^{5,7}

These findings suggest that the N-terminal region of Pol κ contributes to increasing the enzyme processivity.⁶⁰ Finally, recent mutagenesis studies have provided new insights into the role of the DNA binding in enhancing the processivity of DNA polymerases.^{42,59} These results highlight the importance of protein•DNA interactions in modulating the processivity of DNA polymerases and suggest that further studies are needed to fully understand the mechanisms underlying this property.

Classification of DNA and RNA polymerases

Historically, DNA polymerases were first discovered in 1958 by Arthur Kornberg and collaborators, Identifying PolI in *Escherichia coli*.⁶¹ This finding laid the foundation for our understanding of DNA replication. Since then, DNA polymerases have been classified into six major families (A, B, C, D, X, Y, and RT), according to their sequence homology, structural motifs, and biological function.^{7,43} A summary of their classification and representative members is reported in Table 1.2.

Family	Pols	Fidelity	Processivity
A-family	PolI, Pol γ , Pol θ , and Pol ν	High	Moderate
B-family	PolII, Pol α , Pol δ , Pol ϵ , and Pol ζ	High	High
X-family	Pol β , Pol λ , Pol μ , TdT, Pol σ	Low	Low
Y-family	Pol IV, and Pol V, Pol η , Pol ι , Pol κ , Rev1	Low	Low

Table 1.2: **Classification of DNA polymerase families.** The table summarizes the main families of DNA polymerases, their representative members, and their typical fidelity and processivity characteristics. Notably, A- and B-family polymerases are mainly replicative enzymes, and in fact exhibit high fidelity and processivity. By contrast, X- and Y-family polymerases are typically involved in DNA repair and translesion synthesis, so they are characterized by lower fidelity and processivity.^{7,43}

Similarly, RNA polymerases are divided into six main families (A to F) according to their structure, function, and functional role.⁶² In brief:

- Family A: Cellular multi-subunit DNA-dependent RNA polymerases from eubacteria, archaea, and eukaryotes. Includes also chloroplast and vaccinia virus polymerases.
- Family B: Monomeric DNA-dependent RNA polymerases involved in DNA replication, such as primases (Figure 1.2).

- Family C: Viral monomeric DNA-dependent RNA polymerases from bacteriophages (T3, T7, SP6) and yeast mitochondria.
- Family D: Multi-subunit RNA-dependent RNA polymerases (replicases) found in RNA phages such as Q β and MS2.
- Family E: Viral DNA-dependent RNA polymerases encoded by alphaviruses, plant positive-strand RNA viruses, and birnaviruses.
- Family F: Viral RNA-dependent RNA polymerases encoded by picornaviruses, potyviruses, and other segmented RNA viruses.

Analogous functional distinctions are observed in their fidelity and processivity. For instance, RNA Pol II exhibits high transcriptional fidelity, with an error rate of $\sim 10^{-5}$. In contrast, viral RNA polymerases typically have moderate fidelity, ranging from 10^{-3} to 10^{-5} , facilitating rapid adaptation of RNA viruses to various host environments.⁷

A-family

DNA polymerases belonging to the A-family are found across both prokaryotic and eukaryotic organisms and are involved in key processes such as DNA replication and repair.⁴ A-family Pols are generally characterized by a right-hand-like structure, with a palm, fingers, and thumb subdomains, similar to other polymerase families.⁷ However, they also possess unique features that distinguish them from other families, such as the presence of a 5'→3' exonuclease subdomain in PolI and the specific structural adaptations in Pol γ for mitochondrial DNA replication.⁴³ This family includes DNA polymerase I (PolI) in prokaryotes, as well as polymerases γ (Pol γ), θ (Pol θ), and ν (Pol ν) in eukaryotes. PolI, the prototype of this family, was the first DNA polymerase to be discovered, isolated, and structurally characterized.⁶¹ Although it was initially thought to be the primary replicative enzyme in *Escherichia coli*, it is now well established that PolI plays a more specialized role in DNA repair and the processing of Okazaki fragments.² This enzyme exhibits two exonuclease subdomains needed for two different functions: the 3'→5' enables proofreading of misincorporated nucleotides, and the 5'→3' allows for the removal of RNA primers and damaged DNA segments. The coexistence of these multiple enzymatic functions makes PolI a particularly versatile tool in both cellular processes and biotechnological applications.² In eukaryotes, Pol γ is the mitochondrial replicative polymerase, responsible for maintaining the mitochondrial genome.⁶³ Due to the critical nature of mitochondrial DNA, Pol γ is characterized by high fidelity and proofreading activity (exonuclease subdomain 3'→5'), ensuring accurate replication under conditions of oxidative stress and limited DNA repair capacity.⁶⁴ Conversely, Pol θ operates within the nucleus and is mainly associated with DNA repair, particularly in pathways such as translesion synthesis (TLS) and alternative end-joining.^{65–67} Unlike Pol γ , Pol θ lacks proofreading activity and is considered a low-fidelity polymerase, reflecting its ability to bypass DNA lesions and tolerate mismatched bases. Finally, the precise functions of Pol ν are still being elucidated, but like Pol θ , it is thought to be involved in DNA repair mechanisms and to possess low replication fidelity.⁶⁶

B-family

B-family DNA polymerases represent the core of the eukaryotic replication machinery.⁶⁸ These enzymes are structurally conserved and typically composed of multiple subunits. They feature three key domains: the N-terminal region, the exonuclease (Exo) subdomain, and the polymerase (Pol) core.⁴ The function of the N-terminal region remains only partially understood, playing a role in uracil recognition in archaeal homologs but not in eukaryotic counterparts.² The Exo subdomain provides a 3'→5' proofreading function that contributes to the high fidelity of DNA synthesis. The Pol core, adopting a characteristic right-handed architecture, is responsible for the template-guided incorporation of nucleotides.^{4,7} In eukaryotes, four main B-family polymerases carry out distinct functions. Pol α initiates DNA replication and is part of a dual-function complex containing both primase and polymerase activities.⁶⁹ Although essential for laying down RNA/DNA primers, Pol α lacks proofreading capacity due to an inactive exonuclease subdomain. Following initiation, Pol ϵ and Pol δ take over as the main processive polymerases responsible for elongation of the leading and lagging strands, respectively (see Figure 1.2).^{70,71} These enzymes operate with high fidelity, especially in the presence of their accessory factors, and are critical for accurate genome duplication (see Table 1.1). Pol ζ , the fourth B-family member in eukaryotes, diverges from the others in function and fidelity.⁷² Unlike Pol δ and Pol ϵ , it exhibits low accuracy and lacks an active proofreading mechanism.⁴ However, this reduced fidelity is functionally advantageous in the context of translesion synthesis (TLS), where Pol ζ extends DNA synthesis past lesions or mismatches that stall replicative polymerases. This specialized activity makes Pol ζ crucial for DNA damage tolerance and lesion bypass mechanisms.⁷³ Interestingly, B-family polymerases are not exclusive to eukaryotes. Several bacteriophages, such as T4 and Phi29, also rely on family B enzymes as their primary replicases, highlighting the evolutionary conservation and versatility of this polymerase family.^{7,74,75}

X-family

DNA polymerases belonging to family X are small, monomeric enzymes primarily involved in DNA repair, particularly in filling short single-stranded gaps.⁴³ These enzymes are evolutionarily conserved and are found in a wide range of organisms, from bacteria and viruses to yeast and mammals.⁴ This broad conservation highlights their fundamental role in genome maintenance. None of these enzymes possesses a proofreading subdomain, and their intrinsic fidelity of DNA synthesis is much lower than that of a polymerase that functions in cellular DNA replication. The best-characterized member of the family is DNA polymerase β (Pol β), which plays a central role in base excision repair (BER) by inserting the correct nucleotide opposite a damaged base and removing the 5'-deoxyribose phosphate group left by the DNA glycosylase.^{52,76} In addition to Pol β , family X includes Pol λ and Pol μ , as well as terminal deoxynucleotidyl transferase (TdT). Like Pol β , pol λ participates in BER, while like pol μ , it participates in the repair of double-strand DNA breaks and has also been implicated in translesion DNA synthesis (TLS).^{77,78} Another eukaryotic enzyme within the X-family is Pol σ , encoded by the TRF4 gene in *Saccharomyces cerevisiae*. Initially proposed to participate in sister chromatid cohesion, the nature of Pol σ 's enzymatic activity remains controversial.

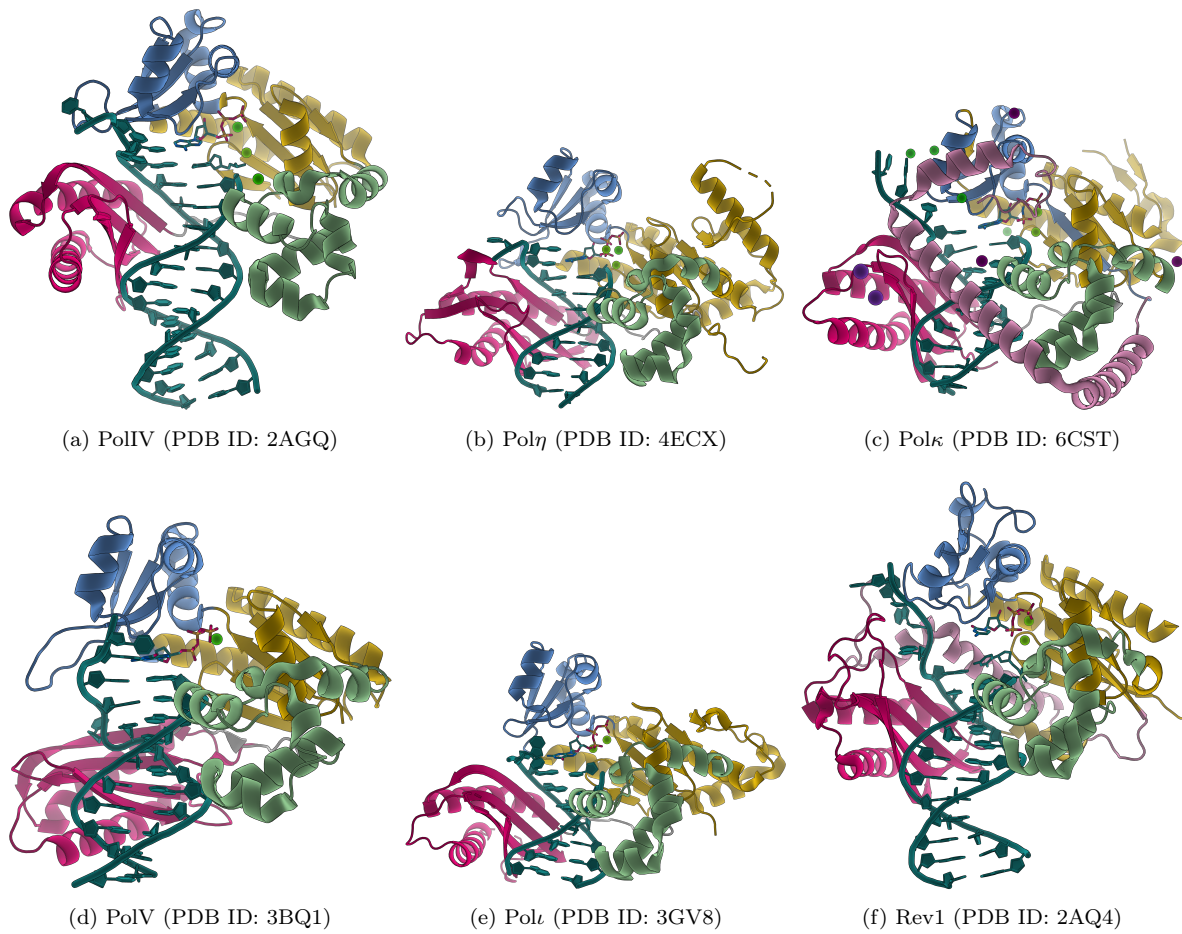


Figure 1.5: **Prokaryotic and eukaryotic Y-family polymerases.**⁴³ Crystal structures of ternary complexes of (a) Archeal PolIV (Dpo4, PDB ID: 2AGQ), (b) Human Pol η (PDB ID: 4ECX), (c) Human Pol κ (PDB code 6CST), (d) Archeal PolV (Dbh, PDB code 3BQ1), (e) Human Pol ι (PDB code 3GV8), and (f) Yeast Rev1 (PDB code 2AQ4) are shown in cartoon. The enzymes are colored to underline their subdomains (palm in gold, finger in lightblue, thumb in lightgreen, and little-finger in magenta). The N-clasp of Pol κ and the N-digit of Rev1 are shown in pink. The incoming nucleotide is drawn in licorice, the Mg²⁺ and K⁺ ions are represented in spheres and colored in green and purple, respectively. Finally, the DNA substrate is in teal cartoon. The figures are made using the PyMOL Molecular Graphics System, Version 3.1.⁴⁰

1.3.3 Y-family: Translesion Synthesis Polymerases

Y-family DNA polymerases are specialized enzymes that play a crucial role in DNA damage tolerance mechanisms.⁵ Their unique structural and functional characteristics enable them to efficiently bypass DNA lesions and maintain genomic stability.⁷⁹ Their discovery emerged in the early 2010s from the realization that certain bacterial proteins, such as UmuC/UmuD' and DinB in *Escherichia coli*, possess DNA polymerase activity.⁴³ This finding paralleled the identification of related enzymes in eukaryotes, most notably polymerase η (Pol η). In humans, mutations in the gene encoding Pol η result in Xeroderma Pigmentosum Variant (XP-V), a hereditary condition marked by extreme sensitivity to UV light and a predisposition to skin cancer.¹³ Unlike many other DNA polymerases, members of the Y-family lack 3'-5' exonuclease activity, and instead contain a distinct structural motif known as the PAD (polymerase-associated domain), also referred to as the wrist or little finger subdomain. Notably, the latter subdomain

does not contain primary sequence motifs that are conserved throughout the Y-family.⁵ In fact, in this subdomain, there are major differences in terms of sequence between Y-family Pols. Yet, this subdomain has a conserved ternary structure, a four-stranded β sheet supported on one side by two long α helices.⁵ This region plays a critical role in shaping substrate specificity and facilitating interaction with distorted DNA structures.⁸⁰ Y-family polymerases are characterized by their inherently low fidelity on undamaged DNA, a trait largely attributed to their more flexible active sites, which can accommodate bulky or irregular DNA lesions.^{81,82} Compared to replicative polymerases, these enzymes establish fewer protein•DNA contacts.²⁷ This is due to the fact that the finger subdomain is smaller than in proteins from other families. Notably, Pol η possess a β -sheet in the LF subdomain (residues 316-324 in PDB ID: 4ECX) that is parallel to the template strand, forming several hydrogen bonds helping to stabilize the complex. In this way, Pol η acts like a molecular splint keeping the damaged DNA in place, facilitating the correct DNA synthesis.³³ It has been experimentally demonstrated that mutating residues within this loop impairs the activity and fidelity of Pol η .⁸³ By contrast, this structural element is not present in other Y family members, which present a gap separating the LF subdomain from the Pol core (Figure 1.3.3).³³ Human Y-family members are primarily involved in translesion synthesis (TLS), a damage tolerance mechanism that enables DNA replication to proceed across damaged sites that would otherwise stall the replication fork (see Figure 1.2 and subsection 1.5.1).⁴³ For example, Pol η is particularly adept at bypassing UV-induced cyclobutane pyrimidine dimers (CPD), inserting the correct nucleotides across the lesion with higher efficiency and accuracy than other polymerases.³³ Furthermore, Y-family polymerases have been implicated in processes such as somatic hypermutation (a mechanism that introduces point mutations into immunoglobulin genes to diversify the antibody repertoire), highlighting their broader relevance in both DNA repair and adaptive immunity.^{84–86}

Structural differences within the family

Due to the diverse chemical properties, structural features, and rigidity of DNA lesions, repair polymerases have evolved distinct structural adaptations that enable them to specialize in bypassing specific types of damage.⁵ These structural differences are mainly present outside the classical Pol domain.⁵¹ Pol κ and Rev1, for example, possess unique N-terminal extensions-termed the N-clasp in Pol κ and the N-digit in Rev1, which contribute to a more enclosed active site architecture (see Figure 1.5c and Figure 1.5f). These structural configurations result in a tighter anchoring of the DNA to the enzyme, potentially enhancing lesion recognition and processivity efficiency (better explained in subsection 1.3.2).⁵ The catalytic core, common to all DNA polymerases (referred to subsection 1.3.1), is connected to the LF subdomain via a disordered and highly flexible region known as the linker.⁵ The length and chemical composition of this linker vary significantly among members of the Y-family, suggesting a possible role in modulating enzyme flexibility and substrate access. Notably, Pol κ , PolIV, and PolV exhibit particularly dynamic linkers that are capable of undergoing substantial conformational rearrangements upon DNA binding.⁸⁷ This flexibility is thought to facilitate the accommodation of bulky DNA lesions, allowing these polymerases to effectively bypass a wide range of structural distortions in the DNA helix.⁵ In contrast, Pol ι have shorter and more rigid linkers, which may limit its ability to accommodate larger lesions but enhance its fidelity on undamaged DNA.^{5,50,57} The structural diversity within the Y-family polymerases reflects their evolutionary adaptation to different DNA damage scenarios, enabling them to perform specialized

roles in maintaining genomic integrity.⁴³

Mind the gap

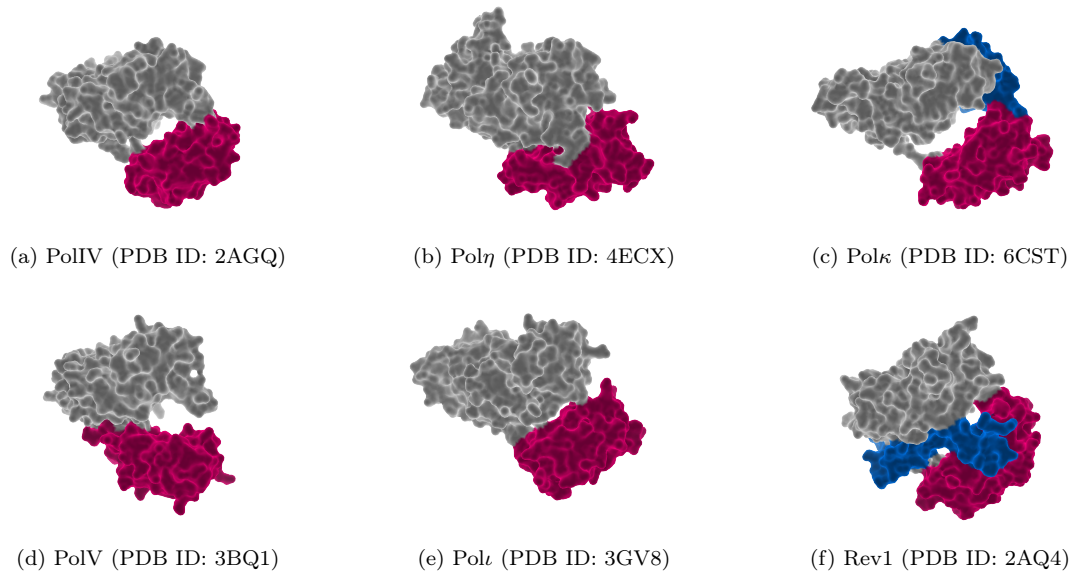


Figure 1.6: Spotlight on the structural gap between the polymerase core and little finger (LF) subdomain in Y-family polymerases.⁵ The crystal structures are the same used in Figure 1.5. The polymerase core (palm, finger, and thumb) is colored in white, while the little-finger subdomain is in magenta, to underline the size of the structural gap between them. The N-clasp of Pol κ and the N-digit of Rev1 are shown in blue. The figures are made using the PyMOL Molecular Graphics System, Version 3.1.3⁴⁰

The conformational changes occurring within the catalytic cycle contribute to the formation of a prominent cleft between the finger and the LF subdomain in Y-family Pols (see Figure 1.6).⁵ The presence and extent of this gap are closely linked to the length and flexibility of the linker region, which typically lacks defined secondary structure and connects the C-terminal end of the thumb subdomain to the LF. A longer and more flexible linker results in a highly mobile LF subdomain.⁵ This large gap is hypothesized to facilitate the accommodation of bulky DNA lesions that would otherwise be sterically hindered. Indeed, the LF subdomain in Pol η rotates away from the catalytic core, probably to open the active site and accommodate bulky lesions.⁸⁸ Notably, polymerases like Pol IV, characterized by a wider gap, exhibit an approximate 14 degrees rotation of the LF between the apo and ternary forms (Figure 1.7). In extreme cases, such as Pol κ and Rev1, the gap is so pronounced that additional structural elements, such as the N-clasp region at the N-terminus, are required to stabilize the enzyme upon DNA binding.⁵

A second key structural element influencing the gap architecture is the length of the loop located at the interface between the fingers and LF subdomains (residue 79-88 in PDB ID: 4ECX). This loop interacts with the template junction, stabilizing it within the gap and helping define the channel through which the junction is positioned.²⁷ The size of the gap directly affects LF mobility and the overall architecture of the polymerase. Notably, recent computational and structural studies have proposed the template junction movement as the rate-limiting step in Pols translocation.^{20,27,90} This suggests that the gap size and LF mobility are critical determinants of the enzyme's translocation efficiency and overall

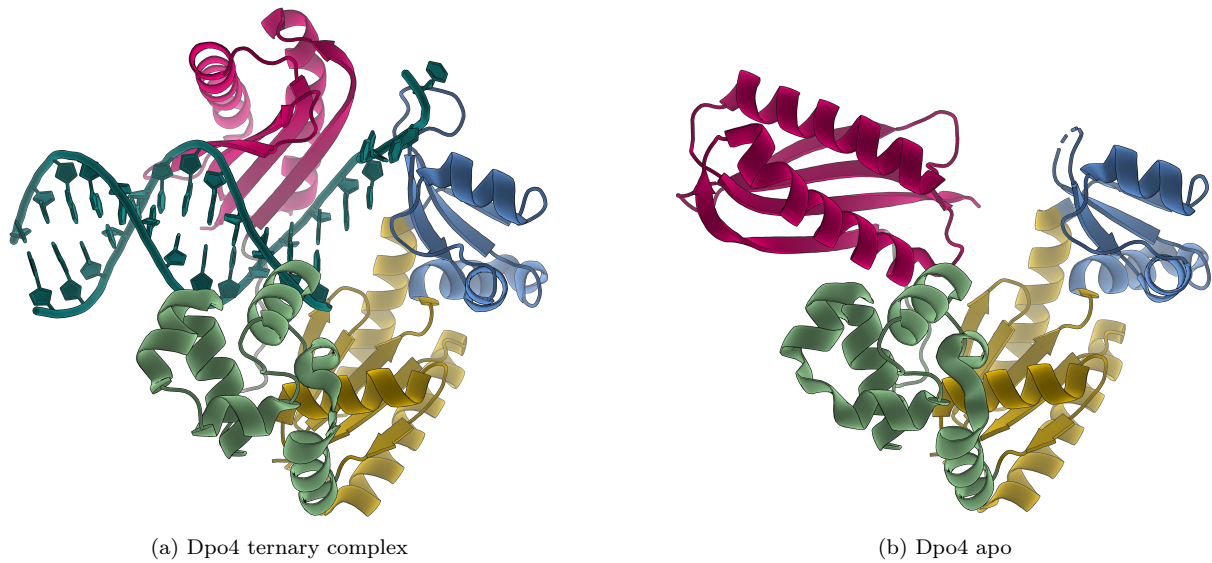


Figure 1.7: **Structures of Dpo4 in apo (PDB ID: 2RDI) and binary complex (Pol•DNA; PDB ID: 2AGQ) forms, respectively.** The little finger subdomain rotates nearly 130 degrees when the enzyme binds to the DNA substrate. By contrast, in *Bacillus* DNA polymerase I (PDB ID: 4BDP; Figure 1.4), this rotation is not observed. In fact, the apo structure and the binary complex are superimposable.^{55, 89}

catalytic activity. Finally, recent studies have proposed a correlation between the DNA-binding capacity of Y-family polymerases and their processivity.⁹¹ Notably, Pols possessing both a highly flexible LF and an expanded gap resulting in higher processivity compared to other Y-family enzymes (see Table 1.1). Nevertheless, the molecular determinants underlying the generally low, yet variable, processivity of Y-family polymerases remain poorly understood, and the direct relationship between structural dynamics and enzymatic function requires further investigation.

1.4 Pols Catalytic Cycle and Translocation Mechanism

The catalytic cycle of DNA polymerases is a complex series of steps, as illustrated in Figure 1.8, that enable the synthesis of new DNA strands by adding nucleotides to the growing chain.⁷ This process is essential for DNA replication, repair, and recombination, ensuring the faithful transmission of genetic information across generations. The cycle can be divided into several key stages, including substrate binding, nucleotide addition, translocation, and product release.²⁷ Each stage is characterized by specific conformational changes in the polymerase structure, which facilitate the precise positioning of the DNA substrate and the incoming nucleotide for catalysis. The catalytic cycle is highly regulated to ensure fidelity and efficiency, with various factors influencing enzyme activity, such as metal ion cofactors, nucleotide availability, and the presence of accessory proteins.²⁷

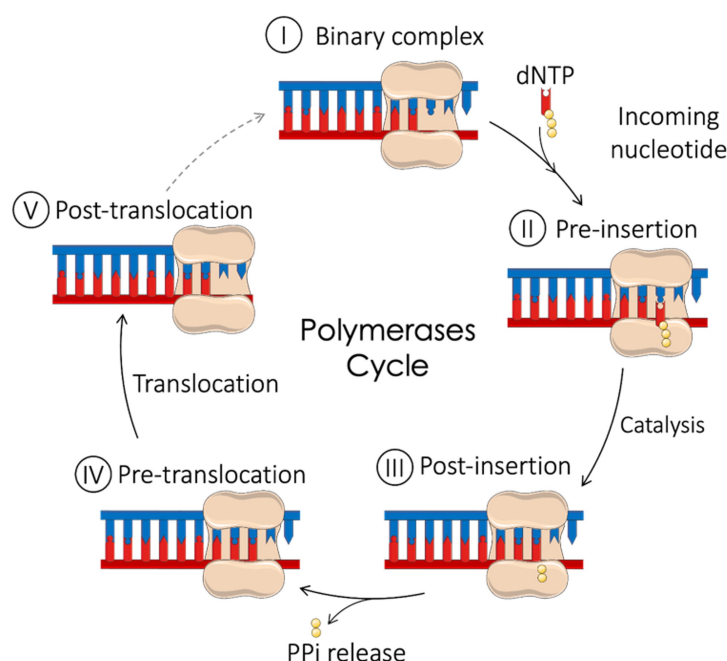


Figure 1.8: **Stepwise catalytic cycle of DNA polymerases.** (I) The enzyme binds to the DNA template and the primer strand, forming a binary complex. (II) The incoming deoxynucleotide triphosphate (dNTP) is positioned opposite the template base. (III) The 3'-OH group of the primer strand attacks the α -phosphate of the dNTP, leading to phosphodiester bond formation (IV) The pyrophosphate (PP_i) is released. (V) The enzyme translocates along the DNA template, moving the DNA double-stranded to the next position for nucleotide addition. After that, The product is released, and the enzyme is ready for another catalytic cycle.²⁷

Nucleotide Addition

The process begins with the binding of the DNA substrate to the polymerase, forming a binary complex (see (I) in Figure 1.8). This initial interaction is crucial for positioning the DNA template strand and the 3' end of the primer strand within the active site of the enzyme. The next step involves the accommodation of an incoming deoxynucleotide triphosphate (dNTP) at the active site, where it is positioned opposite the complementary base on the template strand. ((II) in Figure 1.8) This step can be facilitated by conformational changes occurring in the finger subdomain, which undergoes a closing motion to stabilize the dNTP binding.⁷ During this step, the ternary complex is formed.

Catalysis and Phosphodiester Bond Formation

DNA polymerases catalyze DNA synthesis through a two-metal-ion mechanism, typically involving two magnesium ions (Mg^{2+}) in the active site (see Figure 1.9).^{92,93} This mechanism is also employed by RNA polymerases, as well as by other DNA-binding metalloenzymes such as topoisomerases, ribonucleases, endonucleases and nucleases.^{7,8,94,95} Regardless of their structure or biological function, these enzymes exhibit a remarkable conservation of the active site, which is almost completely preserved across different enzyme families and species.^{7,25,26,28,91,96} The presence of magnesium ions is essential for the catalytic reaction to occur.⁹⁷

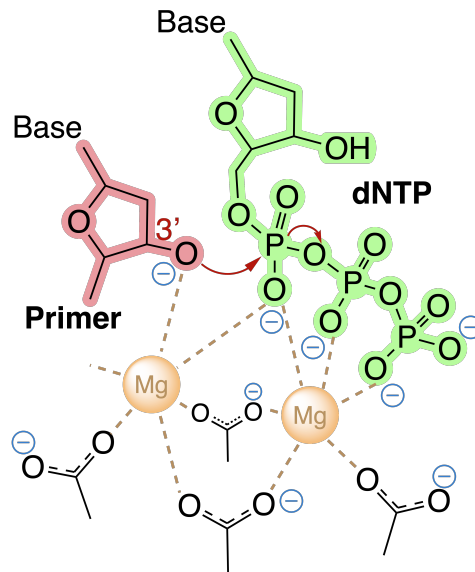


Figure 1.9: **Schematic representation of the catalytic mechanism the two-metal-ion mechanism in DNA and RNA polymerases.** The presence of the two Mg^{2+} ions facilitates nucleotide incorporation through nucleophilic attack of the 3'-OH group of the primer strand (red), leading to phosphodiester bond formation (red arrows). The metal ions also stabilize the pentacovalent transition state, where the α -phosphate of the incoming nucleotide is not yet covalently linked to the 3'-OH group of the growing DNA strand (red), and the pyrophosphate of the incoming nucleotide (green) is detached.

Recent computational and experimental studies have shown that, although polymerases can accommodate manganese ions (Mn^{2+}) in the active site, the enzyme is significantly less efficient in catalyzing the reaction with Mn^{2+} compared to Mg^{2+} .^{98–101} The two magnesium ions, Mg_A and Mg_B , are coordinated by three or four negatively charged residues (Asp or Glu) located in the active site.^{7,26} In addition, Mg_B is also coordinated by the triphosphate moiety of the incoming nucleotide (dNTP). The phosphoryl transfer reaction for nucleotide incorporation is initiated by the deprotonation of the 3'-hydroxy (3'-OH) group of the deoxyribose at the 3' end of the growing strand, leading to the formation of an activated nucleophilic 3'-hydroxide ion (Figure 1.9). This nucleophile then donates a proton to the α -phosphate of the dNTP, forming a pentacovalent phosphate transition state.⁷ Consequently, a new covalent bond is formed between the O3' atom of the primer and the α -phosphate of the incoming nucleotide, which is thereby incorporated into the growing DNA strand (see (III) in Figure 1.8). Simultaneously, the phosphodiester bond between P_α and P_β of the dNTP is cleaved, allowing the release of pyrophosphate (PP_i). The mechanism underlying nucleophile formation in DNA polymerases remains an open question in the field.^{34,96,102} Two main models have been proposed: a protein-mediated and a water-mediated activation mechanism. Warshel and collaborators demonstrated that the DNA polymerase of bacteriophage T7 employs a negatively charged aspartate residue in the active site to trigger deprotonation of the 3'-OH group.¹⁰³ Conversely, a water-mediated mechanism has been observed in two Y-family polymerases (Dpo4 and Pol κ), where a transient bulk water molecule transfers the proton to P_α . Another possible mechanism, proposed by De Vivo and coworkers, is a self-activated model, in which the reaction is initiated by the formation of an intramolecular hydrogen bond between the 3'-OH group and the P_β of the incoming dNTP.⁹⁶ Time-resolved crystallography studies of DNA Pol η have revealed the presence of a third magnesium ion, which some polymerases utilize to neutralize the negative charge of the

transition state and to facilitate protonation of the departing PP_i .^{34,104} In Pol η , this third metal ion is present during the catalytic step and prevents the initiation of translocation.⁹⁶ The presence of this third metal ion has also been observed in other metalloenzymes, such as nucleases, endonucleases, and even in group I introns.^{8,94} Notably, this transient ion acts as an exit shuttle for both PP_i and Mg_B in Pol η , as well as in human exonuclease I.^{94,105} Due to its transient nature, only a few crystallographic structures have captured this third ion.³⁴ In certain metalloenzymes, the release of Mg_B and PP_i initiates DNA translocation, providing the necessary energy for the DNA to shift, leaving free the active site for the next catalytic cycle. Intriguingly, two positively charged residues or monovalent ions occupied the same structural position near the active site in many different metalloenzymes, including DNA/RNA Pols, endonuclease and exonuclease, but also group II intron and spliceosome.^{106–108}

1.4.1 DNA Translocation Mechanism

DNA translocation in Pols is a critical physical step for Pol-mediated nucleic acid polymerization, essential for storing and transmitting genetic information in all living organisms. Once the chemical step has occurred and the PP_i is released ((IV) in Figure 1.8), the newly elongated double-stranded DNA (dsDNA) has to shift along the enzyme to reproduce the initial configuration at the metal-aided reactive center, where nucleotide addition can occur recurrently ((V) in Figure 1.8). Alternatively, the enzyme can unbind the dsDNA instead of translocating along it, thus ending the current cycle.²⁷ High-processive Pols are able to maintain their grip on the DNA template, allowing for continuous synthesis without frequent dissociation and reassociation. By contrast, low-processive Pols may require more frequent translocation events to ensure that the DNA substrate remains bound to the enzyme during synthesis.⁴³ Yet, low-processive Pols are usually specialized in TLS, which means that they are able to bypass DNA lesions, and therefore, they can be less processive than replicative Pols.⁵ The translocation mechanism is a complex process that involves significant conformational changes in the enzyme structure, particularly in the finger subdomain, which plays a crucial role in facilitating the movement of the DNA substrate along the enzyme.²⁷ The finger subdomain is responsible for stabilizing the DNA substrate and the incoming nucleotide, and its conformational changes are essential for the translocation process.^{7,36,37} In many Pols, the finger subdomain undergoes a closing motion upon binding of a correct incoming dNTP and metal ions, which promotes catalysis and DNA translocation. This motion is often coupled with the rotation of a helix, known as the O-helix, which is located in the finger subdomain (see Figure 1.10).^{36,37} This helix is a key structural element that undergoes significant conformational changes during translocation, facilitating the movement of the DNA substrate along the enzyme. The rotation of this helix is thought to be a critical step in the translocation mechanism, allowing the enzyme to shift along the DNA strand and reposition itself for subsequent rounds of nucleotide addition.^{36,37} The O-helix is a conserved structural element found in many DNA polymerases, and its conformational changes are thought to be a key determinant of the translocation mechanism.^{36,37} Thus, the rotation of the O-helix is thought to be uncoupled from the release of PP_i .^{36,37} This separation between the chemical and physical steps of the catalytic cycle is crucial for understanding the translocation mechanism of DNA polymerases, as it allows for the enzyme to efficiently shift along the DNA substrate without being hindered by the release of PP_i .^{36,37} The O-helix is also thought to play a role in the fidelity of DNA synthesis, as its conformational changes can influence the accuracy of nucleotide incorporation and the overall processivity of the enzyme.^{36,37}

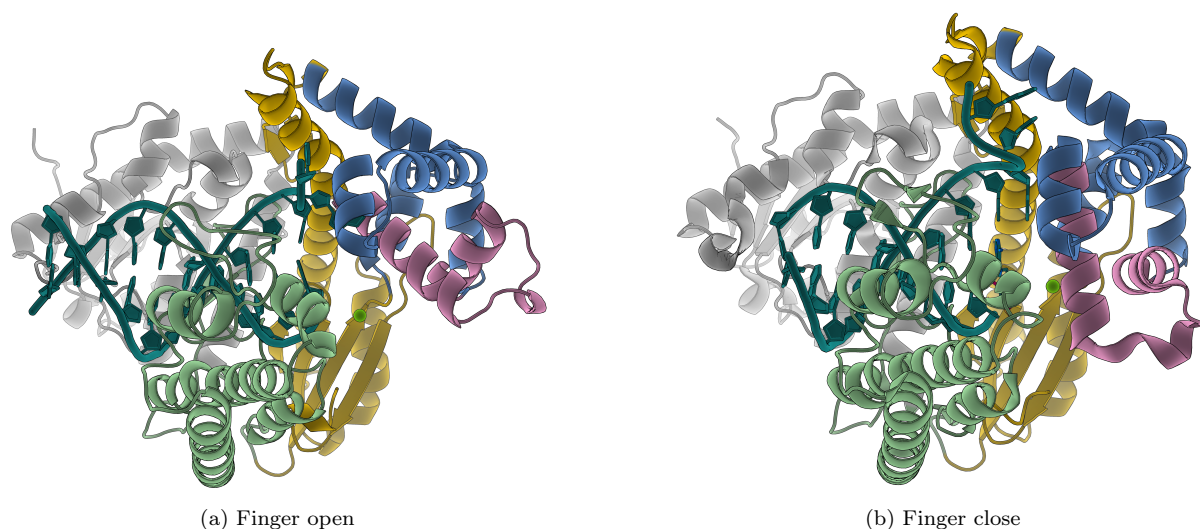


Figure 1.10: **Conformational changes of the fingers subdomain in *Bacillus* DNA polymerases I.** (a) Open state (PDB ID: 4BDP) facilitates incoming nucleotide accommodation. (b) Closed state (PDB ID: 3THV) promotes catalysis and DNA translocation. The O-helix and the two surrounding helices (pink) undergo a closing motion upon binding of a correct incoming dNTP and metal ion (green sphere).

In this regard, two mechanistic models have been proposed to describe Pols translocation, both readapted from models describing motor protein mechanisms.^{37,109} The first is the "power-stroke" model, where the PPI release promotes significant structural changes in the enzyme, which push the DNA substrate away from the active site. Alternatively, the "Brownian-ratchet" model is based on the fast equilibrium between pre- and post-translocation states. In this model, thermal fluctuations, inducing mechanical force and motion, allow the enzyme to shift between initial and final translocation states continuously.¹⁰⁹ These non-equilibrium fluctuations are prevented by the nucleotide binding. The presence of the dNTP in the active site locks the system into the post-translocation state, acting as a "ratchet" stabilizing the system in a given state. Both models are based on the significant conformational motion of a mobile helix located in the finger subdomain. Indeed, the majority of DNA Pols, as well as some RNA Pols follow a translocation mechanism based on the rotation of a helix.^{28,37,75,110–112} Yet, it's not always clear by which of those mechanisms translocation is governed (e.g., chemical or thermal energy). To unravel this, the studies must include other steps of the cycle, such as nucleotide addition.

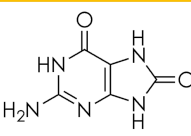
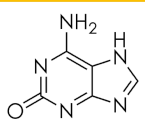
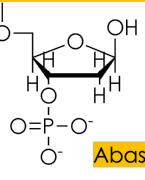
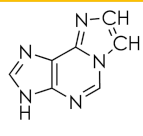
Thus, these two mechanistic models can be applied only to Pols that show significant conformational rearrangements of the overall structure upon translocation, such as replicative polymerases.^{36,66} Notably, Y-family Pols lacks such major structural rearrangements.^{5,27} No mobile structural elements show significant conformation shifts comparing pre- and post-translocation states.^{27,34} To date, no structural evidence supports a rotational movement of the finger subdomain in any Y-family DNA polymerase.^{33,34} Instead, a $\sim 15^\circ$ rotation of the little finger (LF) subdomain has been observed, distinguishing the apo form from the DNA-bound state.^{5,55} Yet, different translocation mechanisms were proposed for PolIV and PolV. PolIV shows a stepwise translocation involving multiple transition states.¹¹³ Conversely, PolV seems to translocate quickly and without the presence of stable intermediate states.¹¹⁴ Pol η and other RdRp show a similar translocation mechanism, which is not based on the rotation of the finger subdomain, but rather on an asynchronous translocation movement.^{20,27,90,115} In these cases, the translocation

mechanism is thought to be split into two steps: the first is the growing strand translocation, while the second is the template strand translocation.²⁷ Recent computational studies also proposed that the presence of a third metal ion lowers the reaction barrier and promotes leaving group departure during catalysis.⁹⁴ However, it remains unclear if such transient binding events can directly affect translocation. For instance, regarding the correlation between ATP binding and hydrolysis to translocation motions in helicases or motor proteins, suggesting the coupling between chemical and physical steps.^{116,117} However, other computational studies prove the absence of coupling between PPi releasing and translocation in RNA polymerase II and in Y-family Pols.^{118,119}

1.5 Specialized DNA Pols and DNA Damage Bypass

1.5.1 Types of DNA Lesions

Most of the replicative polymerases cannot efficiently incorporate the incoming nucleotide in front of template damaged.⁵ Such lesions can be formed spontaneously (endogenous lesions), but also as a consequence of the exposure to various chemical agents (platinum adducts) or radiation (UV-induced lesions).⁷⁹ Consequently, those lesions prevent nucleotide incorporation by replicative polymerases, resulting in the disruption of the catalytic cycle. To overcome these replication blocks, cells use repairing polymerases, able to faithfully incorporate nucleotide opposite to DNA damage. Notably, the majority of repairing polymerases are found in the Y-family, although they are present in all the major families.^{43,79}

Endogenous lesions			
			
8 - oxyguanine	2 - hydroxyadenine	Abasic Site	1,N ⁶ - ethenoadenine
Polη, Polι, Polκ, Polλ, and PolIV	Polλ	Rev1	Poli

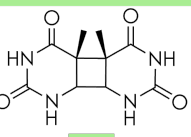
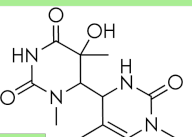
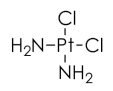
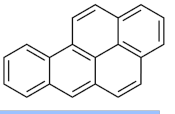
UV-induced		Chemical agents	
			
CPD	6,4 PP	Cisplatin	Benzo(a)pyrene
Polη	Polζ	Polη	Polκ

Figure 1.11: **Types of DNA lesions.** The figure shows the most common types of DNA lesions, including endogenous lesions (deamination, depurination, oxidative damage), UV-induced lesions (cyclobutane pyrimidine dimer and 6,4-pyrimidine-pyrimidone photoproduct), and chemical agents (cisplatin and benzo[a]pyrene). The structures are adapted from.^{79,120,121}

Endogenous lesions

Endogenous DNA lesions arise as a consequence of normal cellular metabolism and include a wide variety of chemical modifications. Common types of damage include deamination, depurination, and oxidative lesions caused by reactive oxygen species (ROS), as well as spontaneous replication errors. Notably, 8-oxoguanine (8-oxoG) is one of the most frequent oxidative lesions and can lead to G→T transversions if not properly repaired.¹²⁰ Other representative examples of endogenous damage, such as 2-hydroxyadenine, abasic sites, and 1,*N*⁶-ethenoadenine, are summarized in Figure 1.11. Although often less dramatic than exogenous damage, these lesions occur continuously and pose a persistent threat to genome stability. Specialized translesion synthesis (TLS) polymerases, including Pol η , Pol κ , and Pol ι , play a critical role in bypassing such damage, enabling replication to proceed but frequently at the cost of fidelity, which can contribute to mutagenesis.

UV-lesions

Exposure to ultraviolet (UV) radiation causes significant DNA damage due to direct absorption of UV photons by DNA bases. This results in photochemical lesions that must be efficiently repaired to maintain genome integrity. The most common UV-induced lesion is the cyclobutane pyrimidine dimer (CPD), formed by covalent bonding between two adjacent pyrimidines, usually thymine-thymine.⁷⁹ CPDs distort the DNA helix, obstructing replication and transcription processes. Another important lesion is the 6,4-pyrimidine-pyrimidone photoproduct (6,4PP), which induces more severe structural distortions.⁷³ These lesions, if unrepaired, can lead to mutations and genomic instability. Specialized DNA polymerases can bypass such damage during replication; for instance, Pol η efficiently bypasses CPDs due to its spacious active site, while bypassing 6,4PPs requires the more flexible Pol ζ polymerase.^{73,89}

Chemical agents

Various chemical compounds can damage DNA, leading to mutations and genome instability. A well-known example is cisplatin, a platinum-based chemotherapeutic widely used in oncology. Cisplatin forms covalent crosslinks between adjacent guanine bases on the same DNA strand, causing substantial distortion of the double helix. These lesions block replication and transcription, ultimately triggering cell death, which underlies cisplatin's cytotoxic effect.¹²² However, cancer cells may develop resistance mechanisms, partly by exploiting translesion synthesis (TLS) pathways that allow DNA replication to continue despite the damage. Another remarkable example is benzo[a]pyrene, a polycyclic aromatic hydrocarbon produced by incomplete combustion processes, including tobacco smoke and grilled foods. Once metabolically activated, it forms bulky adducts with DNA, such as the (+)-trans-BPDE-N²-dG lesion, primarily at guanine residues. These adducts induce strong helical distortions, severely compromising DNA structure and function.¹²¹ If unrepaired, such lesions contribute to mutagenesis and cancer development.¹²³ Like with other bulky DNA lesions, bypass mechanisms involving specialized TLS polymerases may facilitate replication past these obstacles, but often at the cost of increased mutational rates.

1.5.2 Role of Y-family DNA Polymerases in DNA Damage Tolerance

Y-family DNA polymerases are specialized enzymes that facilitate translesion synthesis (TLS), allowing DNA replication to continue past damaged sites that would otherwise stall high-fidelity replicative polymerases. These polymerases are characterized by a more flexible and solvent-accessible active site that can accommodate bulky or distorted DNA lesions.⁵ Although this structural plasticity comes at the cost of reduced fidelity and processivity, Y-family Pols are essential for maintaining genome integrity under genotoxic stress, especially during the S phase. Their ability to bypass specific lesions depends on the unique structural features and mechanisms of each enzyme within the family.

Pol η : Specialized Bypass of UV-Induced DNA Lesions

Pol η is unique among eukaryotic DNA polymerases for its ability to accommodate two unpaired bases from the template strand simultaneously within its active site, enabling efficient and accurate bypass of bulky lesions such as cyclobutane pyrimidine dimers (CPDs) and cisplatin adducts.³³ Structural and biochemical studies have shown that Pol η binds to damaged DNA with similar affinity to undamaged DNA, highlighting its specialized role in tolerating DNA damage.^{33,124} Moreover, it has been proposed that the enzyme inserts four nucleotides before dissociating: two across the lesion and two immediately downstream, after which a loss of stability is observed. Interestingly, Pol η exhibits increased processivity at the lesion site.¹²⁵ Intriguingly, after incorporating the two nucleotides in front of the CPD lesion, Pol η switches to less processive synthesis, and dissociates. This behavior may facilitate the switch to a replicative polymerase after the lesion bypass occurs.¹²⁵ Despite its wide active site capable of accommodating CPDs, Pol η is not well suited to handle lesions that induce more pronounced distortions in the double helix, such as 6,4-pyrimidine-pyrimidone photoproducts (6,4PPs).³³ Efficient bypass of such lesions requires not only a large but also a flexible active site, features lacking in Pol η . In these cases, Pol ζ remains the only polymerase able to accommodate and traverse such severely distorting DNA lesions, due to its broader and more elastic catalytic cleft.⁷³ Key residues, including Arg61 and Gln38, stabilize the lesion-containing DNA, while interactions between the catalytic core and the little finger (LF) subdomain form a molecular splint that preserves the B-form DNA conformation during synthesis.^{33,126} This structural arrangement allows Pol η to insert two adenines across a thymine dimer with near-canonical geometry, making it among the most accurate TLS polymerases for this lesion.¹²⁴ Indeed, its fidelity rate is not affected by the presence of the geometry distortion caused by this lesion.¹²⁷ The LF subdomain also enhances DNA binding and promotes limited processivity during lesion bypass. After incorporating two nucleotides across the lesion and two immediately downstream, Pol η typically dissociates, facilitating polymerase switching to replicative enzymes.^{125,128} Notably, the conserved motif II, located within the finger subdomain (residues 52-73), is essential for Pol η lesion bypass activity. Mutations within this region abolish TLS function, while the Ser62Gly variant has been shown to enhance bypass efficiency for certain lesions, such as 8-oxoG.^{49,129,130} Beyond its canonical role in TLS, Pol η is involved in non-traditional functions, including somatic hypermutation, RNA extension, and chemoresistance mechanisms.^{85,86,104,131} Its ability to tolerate cisplatin-induced lesions can undermine the efficacy of chemotherapeutic treatments, emphasizing the therapeutic relevance of understanding and modulating Pol η activity.¹²²

Xeroderma Pigmentosum Variant (XP-V): The Consequence of Pol η Deficiency

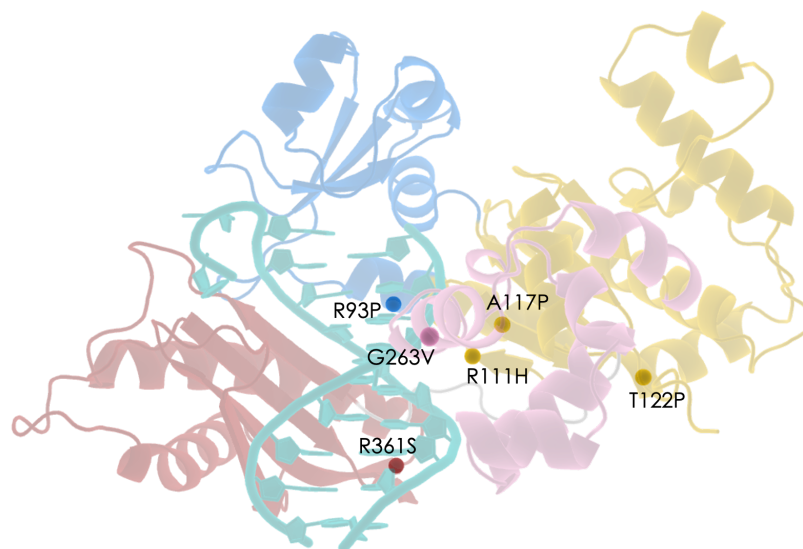


Figure 1.12: **Crystal structure of human Pol η in the presence of XP-V related mutations.** The enzyme (beige) and the DNA substrate (teal) are shown in cartoon representation. The picture underlines the position of the mutation found in XP-V patients, shown in yellow spheres.

Xeroderma pigmentosum variant (XP-V) is a genetic disorder characterized by extreme sensitivity to sunlight and a predisposition to skin cancers. Unlike other XP complementation groups (e.g., XP-A or XP-C), XP-V cells are proficient in nucleotide excision repair (NER) but fail to perform TLS across UV-induced DNA lesions due to Pol η deficiency.¹³ Several missense mutations have been identified within the polymerase core of Pol η in XP-V patients (e.g., Arg93Pro, Arg111His, Ala117Pro, Thr122Pro, Gly263Val, Arg361Ser).^{33, 132, 133} These mutations are hypothesized to induce structural rearrangements or steric clashes that impair polymerase function, although no crystallographic structures of the mutant proteins are currently available. One computational study has proposed that some of these mutations may destabilize the catalytic core or alter DNA binding, but these hypotheses remain experimentally unvalidated.³³ XP-V highlights the critical role of TLS polymerases in genomic maintenance and illustrates how their dysfunction can lead to disease.

Pol κ : Bypass of Bulky Minor Groove DNA Adducts

Pol κ is a Y-family polymerase specialized in bypassing DNA lesions that cause moderate distortions, particularly minor groove adducts such as those induced by benzo[a]pyrene (see Figure 1.11). Its active site strikes a balance between openness and structural constraint, allowing accommodation of bulky adducts with moderate fidelity. Unlike Pol η , Pol κ does not efficiently bypass UV-induced CPDs but excels in processing lesions that require more flexibility in the active site.¹³⁴ The N-terminal region of Pol κ plays a critical role in lesion bypass, as its deletion severely impairs catalytic activity. Additional structural features, such as the N-clasp and the polymerase-associated domain (PAD), contribute to

the stabilization of the DNA substrate and enable lesion discrimination during synthesis. Pol κ exhibits intermediate processivity and often cooperates with other TLS polymerases, such as Pol ζ , to extend DNA synthesis beyond the lesion site (see Table 1.1). Overall, Pol κ acts as an essential player in the coordinated TLS response, allowing DNA replication to proceed past chemically diverse lesions that distort the minor groove, thereby maintaining genomic stability under genotoxic stress.

Pol ι : Unconventional Bypass via Hoogsteen Base Pairing

Pol ι is characterized by its unique ability to utilize Hoogsteen base pairing to bypass DNA lesions that disrupt canonical Watson-Crick geometry, such as 8-oxoguanine and thymine glycol.^{50,135} This polymerase forces the template purine base into a syn conformation, enabling the formation of Hoogsteen pairs with the incoming nucleotide, a mechanism that allows it to replicate across damaged or distorted bases.¹³⁶ Its narrow and rigid active site favors Hoogsteen pairing but also leads to low fidelity, resulting in high error rates even on undamaged DNA.¹³⁷ Despite this, Pol ι plays an important role in DNA damage tolerance by processing lesions that stall replicative polymerases. Structurally, Pol ι differs markedly from other Y-family polymerases, exhibiting a unique DNA binding mode that accommodates the altered base conformations. This specialized activity, while mutagenic, provides a critical backup mechanism to maintain replication continuity in the face of certain lesions.

Rev1: Facilitating Lesion Bypass via Template-Independent Synthesis

Rev1 is a highly specialized Y-family polymerase that operates via a unique mechanism involving an arginine residue (Arg324 in humans) within its active site, which acts as a template surrogate to direct the incorporation of cytosine opposite damaged guanine bases or abasic sites.¹³⁸ During this process, the damaged template base is flipped out of the DNA helix and replaced by the arginine side chain, allowing Rev1 to bypass lesions independently of standard base pairing rules. Beyond its catalytic role, Rev1 functions as a key scaffolding protein within the TLS machinery, interacting with other Y-family and B-family polymerases through its C-terminal region to coordinate lesion bypass and extension steps at stalled replication forks.¹²⁸ This coordination is essential for efficient and accurate TLS, especially for lesions that require a multi-polymerase response.

Chapter II – Materials and Methods

The complementary use of computational and experimental techniques represents the optimal approach for studying complex biological processes, overcoming the limitations of each individual method. Despite the experimental techniques providing a detailed description of the structure and dynamics of biological systems, they are often limited by the time, length, and energy scales that can be probed. On the other hand, computational methods can provide insights into the atomic-level details of biomolecular systems over longer time scales. Still, they are inherently approximate due to the limited computational resources available. Indeed, the time scales of biological processes can range from femtoseconds to seconds, while the length scales can vary from Ångströms to micrometers. This wide range of scales poses a significant challenge for computational methods, which often struggle to capture the full complexity of biological systems. For instance, computational methods can be used to monitor structural changes in proteins, nucleic acids, and lipids, and study the interactions between these biomolecules. These powerful techniques have recently been applied in drug design and discovery, as they can provide insights into the binding mechanisms of small molecules to their targets, the conformational changes of proteins upon ligand binding, and the dynamics of protein-ligand interactions. Depending on the specific biological process under investigation, different computational methods based on varying levels of approximation can be employed (classical mechanics, quantum mechanics, etc.). In this chapter, we will describe the methods used in this work, which include classical molecular dynamics simulations, enhanced sampling methods, and machine learning-based collective variables. We will also discuss the software and tools employed in our research.

2.1 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations are a powerful computational technique used to study the behavior of biomolecules at the atomic level. They allow for exploring the conformational space of biomolecules, providing insights into their structure, dynamics, and interactions. Classical statistical mechanics describes the motion of atoms and molecules, allowing for the simulation of their behavior over time.

2.1.1 Classical Equation of Motion

The theoretical description of molecular systems begins with the fundamental equation of quantum mechanics, the time-dependent Schrödinger equation, which governs the evolution of the quantum state of the system:

$$\mathbf{H}\Psi(\mathbf{r}, \mathbf{R}, t) = i\hbar\partial_t\Psi(\mathbf{r}, \mathbf{R}, t) \quad (2.1)$$

where $\mathbf{H}$ is the Hamiltonian operator, $\Psi(\mathbf{r}, \mathbf{R}, t)$ is the wave function of the system, and $\hbar$ is the reduced Planck constant. The complexity of the Schrödinger equation makes it impossible to solve analytically for systems with more than a few particles. For this reason, Born-Oppenheimer approximation is used to separate the electronic and nuclear degrees of freedom, allowing for the simplification of the equations of motion. In this approximation, the electronic wave function is solved for fixed nuclear positions, and

the resulting potential energy surface is used to describe the motion of the nuclei. In this way, the wavefunction can be rewritten as:

$$\Psi(\mathbf{r}, \mathbf{R}, t) = \psi(\mathbf{R}, t) \cdot \chi(\mathbf{r}, t) \quad (2.2)$$

where $\psi(\mathbf{R}, t)$ is the nuclear wave function and $\chi(\mathbf{r}, t)$ is the electronic wave function. The resulting equation of motion for the nuclei can be expressed as:

$$\mathbf{H}_{\text{nuc}}\psi(\mathbf{R}, t) = i\hbar\partial_t\psi(\mathbf{R}, t) \quad (2.3)$$

where $\mathbf{H}_{\text{nuc}}$ is the nuclear Hamiltonian operator, which is derived from the potential energy surface obtained from the electronic wave function. If we explicitly write the Hamiltonian operator, we can express the equation of motion as:

$$\mathbf{H}_{\text{nuc}} = -\frac{\hbar^2}{2m_i}\nabla_i^2 + U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \quad (2.4)$$

where m_i is the mass of the i -th atom, ∇_i^2 is the Laplacian operator acting on the position of the i -th atom, and $U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ is the potential energy function that depends on the positions of all atoms in the system. The potential energy function is typically derived from quantum mechanical calculations or experimental data and describes the interactions between atoms in the system. The resulting equation of motion for the nuclei can be expressed as:

$$m_i \cdot \frac{d^2\mathbf{r}_i}{dt^2} = -\nabla_i U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \quad (2.5)$$

This equation describes the motion of the nuclei in terms of the forces acting on them, which are derived from the potential energy function. The forces acting on each atom can be expressed as the negative gradient of the potential energy function with respect to the position of that atom. This equation is the basis for classical molecular dynamics (MD) simulations, which are used to study the behavior of biomolecules at the atomic level.

2.1.2 Potential energy and Force fields

To solve the equations of motion for the nuclei (Equation 2.5), we must define a potential energy function that describes the interactions between atoms in the system. The potential energy function is typically derived from quantum mechanical calculations or experimental data and describes the interactions between atoms in the system. The forces acting on each atom can be expressed as the negative gradient of the potential energy function with respect to the position of that atom. This leads to Newton's second law of motion, which can be expressed as:

$$\mathbf{F}_i = m_i \cdot \mathbf{a}_i = m_i \cdot \frac{d^2\mathbf{r}_i}{dt^2} = -\nabla_i U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \quad (2.6)$$

The conservative forces acting on each atom can be expressed as the negative gradient of the potential energy function U , which depends on the positions of all atoms in the system $(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$. Notably, this equation is extremely complicated to solve analytically, especially for large systems. Indeed, classical

molecular mechanics is based on the Bohr-Oppenheimer assumption, which states that the nuclei of atoms are much heavier than the electrons and can be treated as fixed points in space. This simplifies the equations of motion, as the electronic degrees of freedom can be integrated out, leaving only the nuclear motion to be considered. Also, the particles are considered spherically symmetric and charged. The interactions between those particles are mainly based on classical potentials, such as Lennard-Jones or Coulombic potentials. Those interactions are pre-attributed to a specific set of atoms. In this way, the system can be described by a set of classical force fields, which are the union of equations and parameters describing the interactions between atoms. The force fields are used to calculate the potential energy of the system, which is then used to derive the forces acting on each atom.

The classical potential energy function U is typically composed of bonded interactions and non-bonded interaction energies. The total potential energy can be generally expressed as:

$$U(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = U_{\text{bonded}}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) + U_{\text{non-bonded}}(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \quad (2.7)$$

The bonded interactions include bond stretching, angle bending, and dihedral torsions, while the non-bonded interactions include van der Waals and electrostatic interactions. Harmonic potentials typically describe the bonded interactions, while the non-bonded interactions are characterized by Lennard-Jones and Coulombic potentials. The force field parameters are derived from experimental data or quantum mechanical calculations and are used to calculate the potential energy of the system.

$$\begin{aligned} U_{\text{bonded}} = & \sum_i k_{b,i} (r_i - r_{0,i})^2 \\ & + \sum_j k_{\theta,j} (\theta_j - \theta_{0,j})^2 \\ & + \sum_l k_{\phi,l} (1 + \cos(n_l \phi_l - \phi_{0,l})) \\ & + \sum_m k_{\text{im},m} (1 + \cos(n_{\text{im},m} \phi_{\text{im},m} - \phi_{\text{im},0,m})) \end{aligned} \quad (2.8)$$

where $k_{b,i}$, $k_{\theta,j}$, $k_{\phi,l}$, and $k_{\text{im},m}$ are the force constants for bond stretching, angle bending, dihedral torsions, and improper dihedrals, respectively. The parameters $r_{0,i}$, $\theta_{0,j}$, $\phi_{0,l}$, and $\phi_{\text{im},0,m}$ are the equilibrium bond lengths, angles, dihedral angles, and improper dihedral angles, respectively. The non-bonded interactions are typically described by the Lennard-Jones potential and Coulomb's law, which can be expressed as:

$$U_{\text{non-bonded}} = \sum_{i < j} \left[\frac{4\epsilon_{ij}}{12} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \frac{4\epsilon_{ij}}{6} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \sum_{i < j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (2.9)$$

The parameters ϵ_{ij} and σ_{ij} are the well depth and distance at which the potential is zero for the Lennard-Jones potential, while q_i and q_j are the partial charges of atoms i and j . The distance between atoms i and j is denoted as r_{ij} . The first term in Equation 2.9 represents the van der Waals interactions, while the second term represents the electrostatic interactions. The Lennard-Jones potential describes the van der Waals interactions, combining repulsive attractive terms. The repulsive term dominates at short distances, while the attractive term dominates at longer distances. The electrostatic interactions

are described by Coulomb's law, which states that the force between two charged particles is proportional to the product of their charges and inversely proportional to the square of the distance between them. These parameters are typically derived from experimental data or quantum mechanical calculations and are used to calculate the potential energy of the system. Notably, these parameters change depending on the force field used. The choice of force field can significantly affect the results of the MD simulations, and it is essential to carefully select the appropriate force field for the system being studied. Also, classical MD simulations preferably sample low-energy regions of the phase space (because they are more probable), while an ergodic trajectory should also visit high-energy states. Therefore, free energy values and other entropic properties are inaccurate if calculated using classical MD simulations.

Choosing the proper force field

The choice of force field is crucial for the success of molecular dynamics simulations, as it directly impacts the accuracy and reliability of the results. Several factors must be considered when selecting a force field, including the type of biomolecule being studied, the specific properties of interest, and the available computational resources. Indeed, a broad list of Amber force field variants exists, each optimized for different types of simulations and biomolecular systems. For instance, the Amber ff19SB and ff14SB force field is widely used for protein simulations, while the Amber ff15pq force field is designed explicitly for proteins with fluorinated residues.^{139–141} Similarly, the Amber bsc1 force field and OL24 are typically used to treat DNA molecules.^{142–144}

Notably, the accurate simulation of DNA and RNA using molecular dynamics remains a complex challenge, despite significant progress in recent years. Force fields must be sophisticated enough to reproduce the structural and dynamic behavior of these biomolecules, while remaining computationally efficient to allow simulations on biologically relevant timescales. For DNA, the available models are now relatively reliable for describing the canonical double helix under standard conditions. Among the most widely used are the latest versions of the AMBER force fields, such as bsc1 or OL15, which provide good structural stability of the helix and a reasonable reproduction of key parameters such as helical pitch, flexibility, and backbone torsions.^{142,143} However, accurately simulating local deformations, protein recognition, or the formation of alternative structures still requires careful parameter selection and, in some cases, more advanced approaches. The situation is even more complicated for RNA. Unlike DNA, RNA adopts a wide variety of secondary and tertiary structures, often characterized by loops, pseudoknots, and non-canonical interactions, which are particularly difficult to reproduce with classical models. Despite the development of force fields specifically optimized for RNA, such as AMBER OL3 or CHARMM36, RNA simulations remain less reliable than DNA or protein systems.^{145,146} Issues such as structural collapse, loss of stability in loops, or the inability to reproduce complex experimental data are still common. In recent years, more sophisticated approaches have been proposed, including polarizable force fields and machine-learning-based models, which aim to improve the accuracy of nucleic acid simulations. However, these methods typically require significantly higher computational resources and are unsuitable for systematic use in large-scale studies.

Another critical aspect to consider when choosing a force field is the presence of non-standard residues or ions. Indeed, when simulating the system in an explicit solvent, selecting a proper force field to describe water and ions is crucial. For instance, the TIP3P and TIP4P water models are commonly used in MD

simulations, providing a good balance between accuracy and computational efficiency.¹⁴⁷ Similarly, the choice of ion parameters can significantly affect the results of the simulations, especially in systems with high ionic strength or specific ion interactions. The most used ion force fields are the Joung-Cheatham and the Li-Merz.^{148–150} Notably, these force fields include only a subset of the ions commonly found in biological systems, such as sodium, potassium, calcium, and chloride ions. Thus, when simulating systems with other ions, such as magnesium or zinc, it might be necessary to use alternative force fields or develop custom parameters. This is also needed when simulating systems with non-standard residues, as well as ligands and drugs. In such cases, it is often necessary to utilize specialized force fields or to develop custom parameters based on quantum mechanical calculations or experimental data. For instance, the GAFF2 (General Amber Force Field) is a widely used force field for small organic molecules, ligands, and drugs.¹⁵¹ In summary, the choice of force field is a critical step in molecular dynamics simulations, as it directly impacts the accuracy and reliability of the results. It is essential to carefully select the appropriate force field for the system being studied, considering the specific properties of interest and the available computational resources. The choice of force field can significantly affect the results of the MD simulations, and it is important to carefully select the appropriate force field for the system being studied.

2.1.3 Hamiltonian and Lagrangian formalism

The equations of motion can be expressed in terms of the Hamiltonian or Lagrangian formalism. The Hamiltonian formalism is based on the principle of least action, which states that the path taken by a system is the one that minimizes the action, which is defined as the integral of the Lagrangian over time. The Lagrangian is defined as the difference between the kinetic and potential energy of the system:

$$L(\mathbf{q}, \dot{\mathbf{q}}, t) = T(\dot{\mathbf{q}}) - U(\mathbf{q}) \quad (2.10)$$

where $T(\dot{\mathbf{q}})$ is the kinetic energy of the system, $U(\mathbf{q})$ is the potential energy of the system, and $\dot{\mathbf{q}}$ is the velocity of the system. The equations of motion can be derived from the Lagrangian using the Euler-Lagrange equation:

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\mathbf{q}}} \right) - \frac{\partial L}{\partial \mathbf{q}} = 0 \quad (2.11)$$

The Hamiltonian is defined as the Legendre transform of the Lagrangian:

$$H(\mathbf{q}, \mathbf{p}, t) = \dot{\mathbf{q}} \cdot \mathbf{p} - L(\mathbf{q}, \dot{\mathbf{q}}, t) \quad (2.12)$$

where $\mathbf{p}$ is the momentum of the system, defined as $\mathbf{p} = \frac{\partial L}{\partial \dot{\mathbf{q}}}$. The equations of motion can be derived from the Hamiltonian using Hamilton's equations:

$$\dot{\mathbf{q}} = \frac{\partial H}{\partial \mathbf{p}} \quad \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}} \quad (2.13)$$

The Hamiltonian formalism is particularly useful for systems with a large number of degrees of freedom, as it allows for the separation of the equations of motion into two sets of equations, one for the positions and one for the momenta. This separation allows for the use of numerical methods to solve the equations

of motion, such as the Verlet algorithm or the leapfrog algorithm (see subsection 2.1.6).

2.1.4 Statistical ensembles

In molecular dynamics (MD), the trajectory of a molecular system is computed, describing the evolution of atomic positions, velocities, and accelerations as a function of time. This trajectory captures the time-dependent changes in the molecular configuration of the system. MD simulations generate microscopic information, but to connect these data to macroscopic observables, such as energy, pressure, or heat capacity, it is necessary to rely on statistical mechanics. Statistical mechanics describes the behavior of macroscopic systems based on their microscopic properties. During an MD simulation, the system evolves along a trajectory in phase space, governed by the equations of motion and constrained by the chosen thermodynamic conditions defined by a specific statistical ensemble. Each point of this trajectory represents a configuration of the system, specified by the positions and momenta of all particles, consistent with the imposed thermodynamic state. In statistical mechanics, experimentally measurable quantities are expressed as ensemble averages, meaning averages computed over many hypothetical replicas of the system, considered simultaneously. For a classical system at thermodynamic equilibrium, the ensemble average of an observable A is given by an integral over phase space:

$$\langle A \rangle_{ens} = \frac{1}{Z} \int A(\mathbf{p}, \mathbf{q}) \rho(\mathbf{p}, \mathbf{q}) d\mathbf{p} d\mathbf{q} \quad (2.14)$$

where Z is the partition function, $\beta = \frac{1}{k_B T}$, k_B is the Boltzmann constant, and ρ is the probability distribution function in the phase space. In statistical mechanics, different ensembles define the set of accessible microscopic states of a system, depending on the macroscopic thermodynamic conditions imposed. The choice of ensemble determines how physical quantities fluctuate and how averages are computed. In molecular dynamics simulations, these ensembles are mimicked by implementing specific constraints or coupling schemes. The chosen ensemble can be specified in the input file of the molecular dynamics simulation software, and the proper thermostat and barostat algorithms can be used to maintain or reach the desired conditions. The most common ensembles used in molecular dynamics simulations are the canonical ensemble (NVT), the isothermal-isobaric ensemble (NPT), and the microcanonical ensemble (NVE).

Microcanonical ensemble (NVE)

The microcanonical ensemble is used to study systems at constant energy. Still, it is less commonly used in molecular dynamics simulations due to the difficulty of maintaining a continuous energy in a system with numerous degrees of freedom. In this ensemble, the dynamics of the system are governed by the Hamiltonian function, $H(\mathbf{q}, \mathbf{p})$, which represents the total energy of the system as a function of the generalized coordinates $\mathbf{q}$ and momenta $\mathbf{p}$. The Hamiltonian typically takes the form:

$$H(\mathbf{p}, \mathbf{q}) = E = \sum_{i=1}^N \frac{p_i^2}{2m_i} + U(q_1, q_2, \dots, q_N) \quad (2.15)$$

This ensemble is characterized by a fixed number of particles (N), a fixed energy (E), and a fixed volume (V). The probability distribution of the microcanonical ensemble is given by the uniform distribution over the phase space of the system, which is defined as the set of all possible positions and momenta of the particles in the system. The probability distribution is given by:

$$\rho(\mathbf{q}, \mathbf{p}) = \frac{1}{\Omega(E, V, N)} \delta[H(\mathbf{p}, \mathbf{q}) - E] \quad (2.16)$$

where $\Omega(E, V, N)$ is the partition function for the microcanonical ensemble, which is also known as the density of states or phase integral. The partition function for the microcanonical ensemble, also called the density of states or phase integral, is defined as the integral over the phase space of the system, which is the space of all possible positions and momenta of the particles in the system.

$$\Omega(E) = \int \delta[H(\mathbf{p}, \mathbf{q}) - E] d\mathbf{p} d\mathbf{q} \quad (2.17)$$

where the Dirac delta function ensures that only states with energy E contribute to the integral. The density of states is a measure of the number of microstates corresponding to a given energy, volume, and number of particles.

Canonical ensemble (NVT)

The canonical ensemble is characterized by a fixed number of particles (N), a fixed volume (V), and a fixed temperature (T). This ensemble is typically used to study systems at constant temperature. The Boltzmann distribution gives the probability distribution of the canonical ensemble:

$$\rho(\mathbf{q}, \mathbf{p}) = \frac{1}{Z(N, V, T)} e^{-\beta H(\mathbf{q}, \mathbf{p})} \quad (2.18)$$

where $\beta = \frac{1}{k_B T}$, k_B is the Boltzmann constant, and Z is the partition function, which is given by:

$$Z(N, V, T) = \int e^{-\beta H(\mathbf{q}, \mathbf{p})} d\mathbf{p} d\mathbf{q} \quad (2.19)$$

where the integral is taken over the phase space of the system. The partition function is a measure of the number of accessible microstates at a given temperature, volume, and number of particles. The canonical ensemble is frequently employed in molecular dynamics simulations to investigate systems at constant temperature, as it enables the calculation of thermodynamic properties, including free energy, entropy, and heat capacity.

Isothermal-isobaric ensemble (NPT)

The isothermal-isobaric ensemble is characterized by a fixed number of particles (N), a fixed pressure (P), and a fixed temperature (T). This ensemble is used to study systems at constant pressure and temperature. The Gibbs distribution gives the probability distribution of the isothermal-isobaric ensemble:

$$\rho(\mathbf{q}, \mathbf{p}, V) = \frac{1}{\Delta(N, P, T)} e^{-\beta[H(\mathbf{q}, \mathbf{p}, V) + PV]} \quad (2.20)$$

where $\Delta(N, P, T)$ is the partition function for the isothermal-isobaric ensemble, which is given by:

$$\Delta(N, P, T) = \int_0^\infty dV e^{-\beta PV} Z(N, V, T) \quad (2.21)$$

This is an extension of the canonical ensemble, where the volume is allowed to fluctuate.

2.1.5 Ergodicity in MD simulations

In statistical mechanics, the ergodic hypothesis is a fundamental assumption that justifies the equivalence between time averages and ensemble averages of physical observables. Specifically, it states that the trajectory of a single system evolving in time will pass through all accessible microstates compatible with the given macroscopic constraints (such as energy, volume, and number of particles). This implies that the time average of a microscopic observable A along a sufficiently long molecular dynamics trajectory (Equation 2.22) is equal to the ensemble average of the same observable over the ensemble (Equation 2.14).

$$\langle A \rangle_t = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T A(t) dt \quad (2.22)$$

In molecular dynamics simulations, the ergodic hypothesis allows the replacement of ensemble averages by time averages computed along a single trajectory, thus enabling the calculation of macroscopic thermodynamic properties from microscopic dynamics. However, ergodicity is not always guaranteed, especially in complex or glassy systems where the system might get trapped in local minima. Therefore, careful equilibration and sufficiently long simulation times are crucial to achieve representative sampling of the phase space.

2.1.6 Numerical Integration of the equations of motion

Two of the most widely used numerical integrators in classical molecular dynamics simulations are the velocity Verlet and the leapfrog algorithms. Both aim to propagate atomic positions and velocities by discretizing Newton equations of motion over small time steps, but they differ slightly in implementation and properties.

Velocity Verlet Algorithm

The velocity Verlet integrator updates positions and velocities at the same time points, which facilitates straightforward computation of kinetic energy and temperature. The update scheme is:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}(t)\Delta t + \frac{1}{2}\ddot{\mathbf{r}}(t)\Delta t^2 \quad (2.23)$$

$$\dot{\mathbf{r}}(t + \Delta t) = \dot{\mathbf{r}}(t) + \frac{1}{2}[\ddot{\mathbf{r}}(t) + \ddot{\mathbf{r}}(t + \Delta t)]\Delta t \quad (2.24)$$

where $\mathbf{r}(t)$, $\dot{\mathbf{r}}(t)$, and $\ddot{\mathbf{r}}(t)$ are the position, velocity, and acceleration of the particles at time t , respectively. The velocity Verlet algorithm is particularly useful for systems where the forces change slowly over time, as it provides a good balance between accuracy and computational efficiency. It is also straightforward to implement and can be used with various of force fields. However, it can suffer from energy drift over long

simulations, especially in systems with strong forces or high frequencies. This is because the algorithm does not explicitly conserve energy, and numerical errors can accumulate over time, resulting in a gradual drift in the system’s total energy. Applying periodic energy minimization or thermostatics techniques to mitigate this issue and control temperature and energy fluctuations during the simulation is a common approach.

Leapfrog Algorithm

On the other hand, the leapfrog integrator updates positions and velocities at staggered time points. This method is particularly well-suited for Hamiltonian systems and conserves energy more effectively over long simulations. The update scheme is:

$$\dot{\mathbf{r}}\left(t + \frac{\Delta t}{2}\right) = \dot{\mathbf{r}}\left(t - \frac{\Delta t}{2}\right) + \ddot{\mathbf{r}}(t) \Delta t \quad (2.25)$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}\left(t + \frac{\Delta t}{2}\right) \Delta t \quad (2.26)$$

This integrator is well-suited for Hamiltonian systems and generally offers better long-term energy conservation than the velocity Verlet method. It is also efficient, simple to implement, and can be used with various force fields. Nonetheless, like the velocity Verlet method, it may exhibit energy drift during extended simulations, especially in strong interactions or fast motions. To mitigate this, thermostats or regular energy minimization steps are often used to maintain thermal stability and prevent unphysical energy accumulation.

Both velocity Verlet and leapfrog integrators are time-reversible and symplectic, essential properties that help maintain good energy conservation over long molecular dynamics simulations. The velocity Verlet algorithm computes positions and velocities at the same discrete time points, making it easier to calculate kinetic properties, such as temperature. In contrast, the leapfrog method updates velocities at half-integer time steps relative to positions, requiring interpolation to obtain instantaneous velocities; this feature can affect how thermostats are coupled to the system. Regardless of the method, both integrators require a small integration timestep, typically on the order of 1 to 2 femtoseconds, to ensure numerical stability and accurate propagation of the system trajectory.

2.1.7 Temperature and pressure control

The simulated system is ideally isolated in molecular dynamics simulations, evolving with constant total energy (microcanonical ensemble, NVE). However, under most experimental conditions, the surrounding environment maintains a constant temperature and pressure. To better approximate these experimental conditions, it is necessary to use methods that control the system temperature and pressure. Thermostats regulate temperature by simulating heat exchange with an external thermal reservoir, ensuring that the atomic energy distribution corresponds to the canonical ensemble (NVT). Similarly, barostats maintain constant pressure by allowing the simulation box volume to fluctuate, thus simulating the isobaric-isothermal ensemble (NPT). Therefore, using thermostats and barostats enables the reproduction of more realistic thermodynamic conditions, thereby improving the reliability of calculated properties and facilitating comparison with experimental data.

Thermostats

Thermostats are algorithms used to control the temperature of a molecular dynamics simulation by adjusting the velocities of the particles. There are several types of thermostats, including:

- **Berendsen thermostat:** This method couples the system to a heat bath, allowing for gradual temperature adjustments. It is simple to implement, but it can lead to non-physical behavior if not used carefully.¹⁵²
- **Langevin thermostat:** This method introduces a friction term to the equations of motion, simulating the effect of a heat bath. It is widely used in molecular dynamics simulations and can be more stable than the Berendsen thermostat.¹⁵³
- **Nosé-Hoover thermostat:** This method introduces an additional degree of freedom to the system, allowing for more accurate sampling of the canonical ensemble. It is widely used in molecular dynamics simulations but can be more complex to implement.^{154, 155}
- **Andersen thermostat:** This method randomly re-samples particle velocities from a Maxwell-Boltzmann distribution at regular intervals, effectively simulating collisions with a heat bath. It is simple to implement but can disrupt the dynamics of the system.¹⁵⁶
- **Velocity rescaling thermostat:** This method adjusts the velocities of the particles to match the desired temperature. While it is simple to implement, it can lead to non-physical behavior if not applied carefully, as it may distort the natural distribution of velocities and disrupt the correct sampling of the canonical ensemble.¹⁵⁷

Barostats

Barostats are algorithms used to control the pressure of a molecular dynamics simulation by adjusting the volume of the simulation box. There are several types of barostats, including:

- **Berendsen barostat:** Couples the system to an external pressure bath, enabling smooth volume adjustments. Although easy to implement, it does not rigorously reproduce the isothermal-isobaric ensemble and may yield non-physical results if used inappropriately.¹⁵²
- **Parrinello-Rahman barostat:** Introduces additional degrees of freedom, allowing anisotropic scaling of the simulation box for accurate sampling of the NPT ensemble. It is widely used in practice, though its implementation is more complex.¹⁵⁸
- **Monte Carlo barostat:** Employs stochastic volume changes using Monte Carlo sampling to mimic pressure fluctuations. While conceptually straightforward, it is generally less efficient than deterministic methods.¹⁵⁹

2.1.8 Explicit solvent and PBC

In molecular dynamics simulations, systems are typically composed of a finite number of particles. However, to accurately mimic the bulk properties of materials or biomolecules in solution, it is essential to

minimize finite-size effects and surface artifacts. This is achieved by applying periodic boundary conditions (PBC), which conceptually replicate the simulation box infinitely in all spatial directions. Under PBC, when a particle moves out of the primary simulation box, it re-enters from the opposite side, maintaining a constant density and avoiding artificial boundaries. This approach allows the system to emulate an infinite system, ensuring more realistic thermodynamic and structural properties. Implementing PBC requires careful treatment of long-range interactions, often handled by methods such as Particle Mesh Ewald (PME) for electrostatics, ensuring proper energy and force calculations.^{147, 160, 161}

2.2 Sampling the Rare Events

Although classical molecular dynamics techniques are a powerful tool for studying the dynamics of biomolecular systems, they are often limited by the timescales that can be simulated. Many biological processes, such as protein folding, ligand binding, and conformational changes, occur on timescales that are orders of magnitude longer than what can be achieved with standard MD simulations. This limitation arises from the high energy barriers and complex free energy landscapes that characterize these processes, which lead to rare events that are difficult to sample adequately. As a result, classical MD simulations often struggle to sample rare but important events occurring on timescales beyond what is computationally feasible. In rare events, the free energy barrier between reagents and product is typically higher than the thermal energy $k_B T$, leading to a low probability of observing the transition in a standard MD simulation. To overcome this limitation, the so-called enhanced sampling methods have been developed to accelerate the exploration of configurational space and improve the sampling of rare events.

2.2.1 Enhanced Sampling Methods

Starting with the pivotal work of Torrie and Valleau,¹⁶² enhanced sampling methods have been developed to improve the sampling of rare events in molecular dynamics simulations. These methods aim to accelerate the exploration of configurational space by modifying the potential energy surface or introducing bias potentials that enhance the sampling of specific regions of the phase space. Doing so allows for more efficient sampling of rare events and facilitates the calculation of free energy differences, reaction rates, and other thermodynamic properties. Enhanced sampling methods can be broadly classified into static and dynamic methods. Static methods modify the potential energy surface to enhance sampling, while dynamic methods introduce bias potentials that evolve to guide the system through rarely visited states.

2.2.2 Towards collective variable space

In natural science, praxis describes the problem in terms of a few subsets of relevant variables. Those variables are typically referred to as order parameters in physics, while they are known as reaction coordinates in chemistry and biology. In any case, these variables are needed to reduce the dimensionality of the problem and make it more feasible in terms of theoretical and computational resources. Those CVs are functions of coordinates ($CV \equiv s(\mathbf{r})$) and are typically chosen to capture the essential features of the system dynamics. In the majority of the enhanced sampling algorithms, an external bias potential

$U(s(\mathbf{r}))$, a function of those CVs, is added to the system Hamiltonian to enhance the sampling of the configurational space. The idea is to increase the frequency of the slow modes of the system, allowing the system to overcome the free energy barriers and explore the configurational space more efficiently. The bias potential changes according to the enhanced sampling method used. Regardless of the algorithm involved, choosing the proper CVs set is crucial for the success of the enhanced sampling process (as discussed in section 2.3).

In the framework of enhanced sampling techniques, we typically aim to reconstruct the equilibrium properties of the system as they would be observed in the canonical ensemble, where temperature is kept constant through the action of a thermostat. In this context, the microscopic configurations $\mathbf{r}$ follow the Boltzmann distribution (Equation 2.18), which describes the probability of finding the system in a particular configuration at a given temperature. Indeed, the equilibrium probability distribution of the system is given by:

$$\rho_{equil}(\mathbf{r}) = \int d\mathbf{s} \rho(\mathbf{r}) \delta[\mathbf{s} - \mathbf{s}(\mathbf{r})] \quad (2.27)$$

where $\rho(\mathbf{r})$ is the canonical probability distribution (Equation 2.18) and $\delta[\mathbf{s} - \mathbf{s}(\mathbf{r})]$ is the Dirac delta function that ensures that only configurations with the desired collective variables $\mathbf{s}$ contribute to the integral. This equation describes the equilibrium distribution of the system in terms of the collective variables, which are functions of the atomic coordinates $\mathbf{r}$. Using the statistical mechanical formalism, we can express the free energy of the system (FES) in the CVs space:

$$F(\mathbf{s}) = -k_B T \ln \rho_{equil}(\mathbf{s}) + c \quad (2.28)$$

where c is a constant that can be chosen to set the free energy to zero at a reference point. This reduced-free energy is usually smoother than the original one, which reflects the rough phase space of such complex systems.

Static Methods

Static methods modify the potential energy surface to enhance sampling. These methods include umbrella sampling, metadynamics, and replica exchange molecular dynamics (REMD).^{162–165} In all of them, an external bias potential $U(s(\mathbf{r}))$ is added to the system's Hamiltonian, accelerating the dynamics. This bias potential is time-independent and is a function of the CVs. The presence of this bias modifies the energy landscape of the system. This modification results in a different probability distribution:

$$P(s) = \int d\mathbf{r} \rho(\mathbf{r}) \delta[s - s(\mathbf{r})] \quad (2.29)$$

where $\rho(\mathbf{r})$ is the canonical probability distribution (Equation 2.18). Combining Equation 2.29 with Equation 2.28, we can express the free energy in the CVs space as:

$$F(\mathbf{s}) = -k_B T \ln h(\mathbf{s}) - U(\mathbf{s}) = F_{bias}(\mathbf{s}) - U(\mathbf{s}) \quad (2.30)$$

where $h(\mathbf{s})$ is the normalized histogram of the data collected during the simulation, and $F_{bias}(\mathbf{s})$ is the non-physical free energy associated with the bias potential. With this relationship, we can reconstruct

the unbiased free energy landscape of the system $F(\mathbf{s})$ from the biased free energy $F_{bias}(\mathbf{s})$ and the added bias potential $U(\mathbf{s})$. The recovery of unbiased free energy from biased simulations is a crucial step, known as reweighting.

$$\rho_{equil}(\mathbf{r}) \propto \rho_{bias}(\mathbf{r})e^{\beta U(\mathbf{s}(\mathbf{r}))} \quad (2.31)$$

Notably, the reweighting procedure is valid only if the bias potential is time-independent, which is the case for static methods, such as umbrella sampling. More recent methods, such as metadynamics and OPES,^{163,166} introduce a time-dependent bias potential that evolves during the simulation. In this case, the reweighting procedure is more complex and requires additional considerations to account for the time-dependent nature of the bias potential.

Time-dependent methods

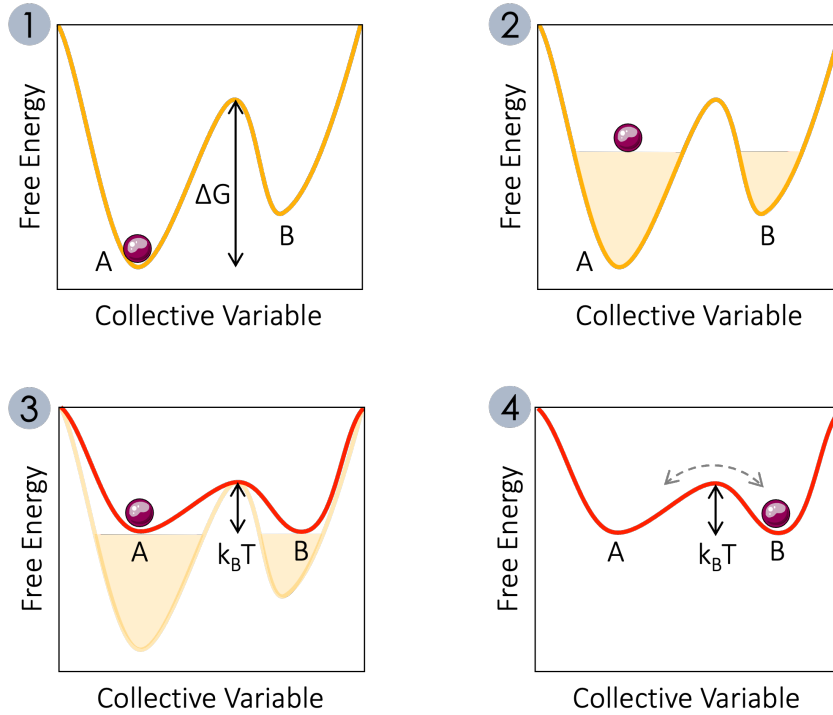


Figure 2.1: **Schematic representation of enhanced sampling methods.** (1) These methods are crucial for addressing rare events in molecular dynamics simulations, where the system, represented by the purple ball, may remain trapped in a local minimum (A) and must overcome a free energy barrier ΔG to explore the configurational space and reach another minimum (B). (2) To facilitate exploration, a time-dependent bias potential $U(\mathbf{s}, t)$ (yellow), defined as a function of collective variables (CVs), is gradually introduced. (3) The bias is continuously updated based on the system's dynamics, enabling efficient sampling of the configurational space. As the total potential (red line), given by the sum of the original potential (yellow line) and the applied bias (yellow surface), lowers the effective energy barriers to the order of the thermal energy $k_B T$, transitions between states become more likely. (4) As a result, the system can efficiently explore otherwise inaccessible regions of the free energy landscape.

Time-dependent methods introduce a bias potential that evolves during the simulation, allowing for more efficient sampling of rare events. These methods include metadynamics, on-the-fly probability en-

hanced sampling (OPES),¹⁶⁶ and adaptive biasing force (ABF). In these methods, the bias potential $U(\mathbf{s}, t)$ is a function of the CVs and evolves over time, allowing the system to explore the configurational space more efficiently. The time-dependent nature of the bias potential allows for a more efficient exploration of the configurational space, as the bias potential is continuously updated based on the system's dynamics. Notably, applying a time-dependent bias potential makes it still possible to recover the unbiased observables of the system. Thus, the reweighting procedure is possible under the condition that, at every instant t_i , the system is in equilibrium with the instantaneous bias potential $U(\mathbf{r}) + U(\mathbf{s}, t_i)$. This condition is satisfied if the bias potential is updated slowly compared to the system's dynamics. This allows the system to equilibrate with the instantaneous bias potential at each time step and converge to a constant value.

For instance, let's consider the case of well-tempered metadynamics, where the bias potential is built up over time by adding Gaussian functions centered on the CVs. The bias potential at time t is given by:

$$U_n(\mathbf{s}) = \sum_{i=1}^N \mathbf{G}(\mathbf{s}, \mathbf{s}_i) e^{\frac{\beta U_{i-1}(\mathbf{s}_i)}{\gamma-1}} \quad (2.32)$$

where $\mathbf{G}(\mathbf{s}, \mathbf{s}_i)$ is a Gaussian function centered in $\mathbf{s}_i$, $U_{i-1}(\mathbf{s}_i)$ is the bias potential at the previous time step, and $\gamma > 1$ is a parameter called bias factor that controls the spread of the bias probability distribution ρ_{bias} . Here, the reweighting procedure is modified as follows:

$$\rho_{equil}(\mathbf{r}) = \rho_{bias}(\mathbf{r}) \cdot e^{\beta U_n(\mathbf{s}, t) - c(t)} \quad (2.33)$$

where $\rho_{bias}(\mathbf{s})$ is the biased probability distribution and $c(t)$ is a normalization constant that ensures the probability distribution is normalized correctly at each time step.

$$\rho_{bias}(\mathbf{s}) \propto \frac{[\rho_{equil}(\mathbf{s})]^{\frac{1}{\gamma}}}{\int d\mathbf{s} [\rho_{equil}(\mathbf{s})]^{\frac{1}{\gamma}}} \quad (2.34)$$

$$c(t) = -k_B T \ln \frac{\int d\mathbf{s} \cdot e^{-\beta F(\mathbf{s})}}{\int d\mathbf{s} \cdot e^{-\beta F(\mathbf{s}) + U_n(\mathbf{s}, t)}} \quad (2.35)$$

2.2.3 On-the-fly Probability Enhanced Sampling (OPES)

On-the-fly Probability Enhanced Sampling (OPES) is a powerful enhanced sampling method that combines the advantages of static and time-dependent methods.¹⁶⁶ It allows for efficient exploration of configurational space by continuously updating the bias potential based on the system's dynamics, while recovering unbiased observables through reweighting. OPES is particularly useful for systems with complex free energy landscapes, where traditional methods often struggle to effectively sample rare events. The revolutionary idea behind OPES is using of the probability distribution to build the bias potential, and not the energy of the system as in other time-dependent bias methods. Indeed, we can define a target probability distribution $\rho_{tg}(\mathbf{s})$ in the CVs space. This is possible by building a bias potential $U(\mathbf{s})$ that

is proportional to the logarithm of the Boltzmann probability distribution $\rho(\mathbf{s})$:

$$U(\mathbf{s}) = -k_B T \ln \frac{\rho(\mathbf{s})}{\rho_{tg}(\mathbf{s})} + c \quad (2.36)$$

This bias function is designed to quickly become quasi-static, allowing the system to explore the configurational space efficiently. Consequently, the physical properties, such as free energy, can be accurately recovered through classical reweighting (Equation 2.31).

Depending on the choice of the target probability distribution $\rho_{tg}(\mathbf{s})$, we can define different variants of OPES.

Uniform variant (OPES)

The original OPES method uses a uniform target distribution (Equation 2.34), which allows for efficient exploration of the configurational space without biasing the system towards specific regions.¹⁶⁷ This variant is particularly useful for systems with complex free energy landscapes, where traditional methods may struggle to sample rare events effectively. In this framework, the bias potential at step n is defined as:

$$U(\mathbf{s})_n = \left(1 - \frac{1}{\gamma}\right) k_B T \ln \left(\frac{\rho_n^{equil}(\mathbf{s})}{Z_n} + \varepsilon \right) \quad (2.37)$$

Explore variant (OPES EXPLORE)

The OPES EXPLORE variant is designed to enhance the sampling of specific regions of the configurational space by using a target distribution that is biased towards those regions. This variant is particularly useful for systems where specific conformational states or transitions are of interest, such as protein-ligand binding or conformational changes in biomolecules.¹⁶⁸ The bias potential in this case is defined as:

$$U(\mathbf{s})_n = (\gamma - 1) k_B T \ln \left(\frac{\rho_n^{WT}(\mathbf{s})}{Z_n} + \varepsilon \right) \quad (2.38)$$

where $\rho_n^{WT}(\mathbf{s})$ is the well-tempered target distribution sample during the simulation:

$$\rho_n^{WT}(\mathbf{s}) = \frac{1}{n} \sum_i^n \mathbf{G}(\mathbf{s}, \mathbf{s}_i) \quad (2.39)$$

2.3 Machine learning based collective variables

In recent years, machine learning (ML) techniques have gained increasing attention in the field of molecular dynamics (MD) simulations, particularly in the context of enhanced sampling methods.²³ Traditional MD simulations are often limited by the timescales they can access, making sampling rare events or slow conformational transitions extremely challenging. ML-based collective variables (CVs) offer a promising solution by providing data-driven representations that capture the essential dynamical features of complex biomolecular systems.¹⁶⁹ Unlike conventional CVs, such as distances, angles, or torsions, which are typically selected based on chemical intuition or prior knowledge, ML-based CVs are derived from the analysis of the simulation data itself. These variables are trained to encode relevant information about

the system’s slow modes or metastable transitions, thus enabling a more efficient exploration of the conformational space. As a result, ML-based CVs facilitate the calculation of free energy surfaces (FES), reaction rates, transition pathways, and other thermodynamic properties with significantly improved accuracy and efficiency.

This section provides an overview of the main types of machine learning methods applied to constructing collective variables, focusing on supervised, unsupervised, and semi-supervised learning paradigms and their relevance in enhanced sampling.

2.3.1 Different classes of Machine Learning methods

Machine learning encompasses a variety of techniques, each suited to specific types of problems. The following classes of ML methods are particularly relevant to the development of collective variables in molecular simulations:

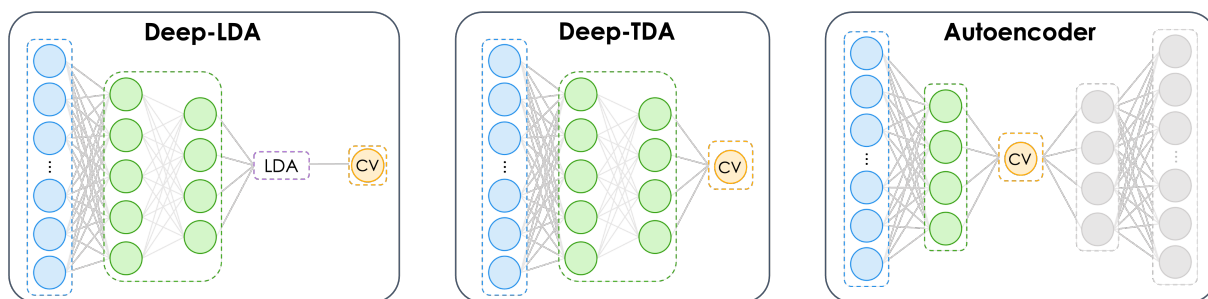


Figure 2.2: **Overview of different neural network architectures.** All of them are based on the same principle of learning a low-dimensional representation of the input data or descriptors (blue) through neural networks (green), but they differ in how they process the data and obtain the output. The output can be either a set of CVs (yellow). (a) Deep-LDA, which learns low-dimensional CVs by optimizing separation between labeled metastable states.¹⁷⁰ Different methods use this architecture employing different statistical methods to obtain the output (here we used LDA as statistical method). The CVs are obtained as linear combinations of output. (b) Deep-TDA, which incorporates physically motivated targets and constraints to learn expressive CVs.¹⁷¹ By contrast to architecture (a), here the outputs of the neural network are directly used as CVs, i.e., without applying any statistical method. (c) Autoencoders, which compress high-dimensional input configurations into a latent space that serves as effective CVs, and a decoder that reconstructs the original input from the latent space.¹⁷²

1. **Deep Learning:** Deep learning approaches, especially deep neural networks (DNNs), are capable of learning complex, nonlinear relationships between input data and target outputs (represented in Figure 2.2). In the context of MD, DNNs can be trained to extract latent features from molecular configurations, enabling the construction of CVs that capture the slow degrees of freedom of the system. Autoencoders and variational autoencoders (VAEs) have been employed to compress high-dimensional structural data into a low-dimensional latent space that serves as a compact and informative CV.^{27, 172–174} These methods can also be used to learn the free energy landscape directly from the data, providing a powerful tool for enhanced sampling.
2. **Kernel Methods:** Kernel-based models, such as Support Vector Machines (SVMs) and Gaussian Processes (GPs),^{175, 176} are effective for learning from small to moderate-sized datasets. These methods work by projecting the input data into a high-dimensional feature space where linear

separation or regression becomes easier. In MD, GPs have been used to construct free energy surfaces from sparse data, incorporating uncertainty quantification and smooth interpolation between sampled states.

3. **Ensemble Methods:** Ensemble methods, including random forests¹⁷⁷ and gradient boosting machines,¹⁷⁸ aggregate multiple base learners to improve predictive performance. These approaches offer robustness and interpretability and can be trained on structural descriptors to classify conformational states or transitions. They are particularly useful when dealing with noisy or heterogeneous data from MD simulations.
4. **Dimensionality Reduction:** Dimensionality reduction techniques, such as Principal Component Analysis (PCA),¹⁷⁹ time-lagged Independent Component Analysis (TICA),¹⁸⁰ and t-distributed Stochastic Neighbor Embedding (t-SNE),¹⁸¹ aim to reduce the number of degrees of freedom in the system while preserving the essential dynamic or structural features. TICA, in particular, identifies the slowest modes of motion in time-series data and has become a popular method for defining CVs in Markov State Models (MSMs) and enhanced sampling techniques.¹⁸²

2.3.2 Data-driven approaches to CV design

A data-driven approach to CV design leverages the large volumes of trajectory data generated from MD simulations to automatically identify relevant patterns and features. This approach eliminates the need for manual selection of descriptors, allowing for the direct extraction of system-specific information from the data. It can reveal hidden dynamics and correlations that may be difficult to capture using conventional CVs. By integrating ML into the analysis workflow, researchers can define CVs optimized for the dynamical processes of interest, improving the performance of enhanced sampling methods such as metadynamics, umbrella sampling, and transition path sampling.¹⁸³ Below, we explore how different learning paradigms, e.g., supervised, unsupervised, and semi-supervised, can be applied in this context.^{172, 173}

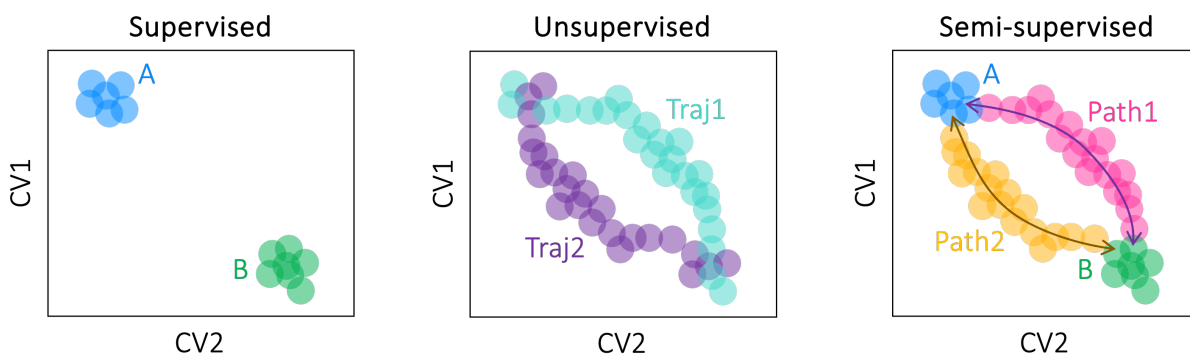


Figure 2.3: **Overview of different data-driven approaches for collective variable (CV) design in molecular dynamics simulations.** The figure schematically illustrates three paradigms: (left) Supervised learning, to train a classifier CV from classical MD data; (middle) Unsupervised learning, to discover new metastable or transition states from multiple trajectories data; and (right) Semi-supervised learning, to integrate both approaches in a multitask framework, employing classification and discovery to improve CV quality and generalization.¹⁷²

Supervised learning

Supervised learning involves training a model on labeled data, where each configuration is associated with a known label, such as a metastable state or a value along a reaction coordinate. In molecular dynamics, supervised models are commonly used to classify conformations (e.g., active vs. inactive protein states) or to regress progress variables related to conformational transitions. The output of supervised classifiers can be interpreted as continuous reaction coordinates and used directly as collective variables (CVs) in enhanced sampling simulations. Recent deep learning-based techniques, such as Deep Linear Discriminant Analysis (Deep-LDA) and Deep Targeted Discriminant Analysis (Deep-TDA), have further advanced this approach (see Figure 2.2).^{170,171} These methods learn low-dimensional CVs by optimizing the separation between labeled metastable states (Deep-LDA) or by incorporating physically motivated targets and constraints (Deep-TDA). For instance, labeled data collected from classical MD simulations can be used to train such models (see left panel in Figure 2.3). Once trained, the model can be used to enhance sampling along the supervised CVs, leading to multiple transitions between metastable states A and B. Notably, in this supervised scenario, the model can effectively capture biologically relevant features associated with states A and B, while remaining agnostic to other, unexplored regions of the phase space.

Unsupervised learning

Unsupervised methods such as clustering (e.g., k-means, DBSCAN, hierarchical clustering) and dimensionality reduction techniques (e.g., Principal Component Analysis (PCA) and Time-lagged Independent Component Analysis (TICA)) can be applied to group similar molecular configurations or to identify low-dimensional manifolds that capture the essential dynamical features of the system.^{179,180,184} These approaches are particularly valuable for unbiased identification of metastable states and dominant transition pathways. For example, TICA is widely used to extract the slowest collective motions from simulation data, which can subsequently be employed as CVs for constructing Markov State Models (MSMs) or for guiding enhanced sampling simulations.^{182,183} Another prominent unsupervised approach is the use of autoencoders, deep neural networks trained to learn low-dimensional representations by compressing high-dimensional molecular configurations into a latent space (see Figure 2.3). The resulting latent variables can serve as effective CVs for enhanced sampling. Notably, autoencoders are often integrated into iterative workflows that alternate between training and sampling. In this scheme, the model is trained on the configurations generated by previous simulations, and new simulations are launched using the learned representation as a biasing variable. This iterative refinement process leads to increasingly accurate and system-specific CVs, ultimately improving the sampling efficiency in subsequent simulation rounds.¹⁷³ As an illustrative example, two independent trajectories sampled from molecular dynamics simulations are projected onto a two-dimensional space of collective variables (CVs) obtained via unsupervised dimensionality reduction. Although the data are unlabeled, it is possible to identify distinct dynamical behaviors: the trajectories follow different pathways across the CV space, suggesting the presence of multiple transition routes between metastable regions. Such low-dimensional embeddings are useful for revealing hidden kinetic features of the system, even in the absence of predefined state labels (see central panel of Figure 2.3).

Semi-supervised learning

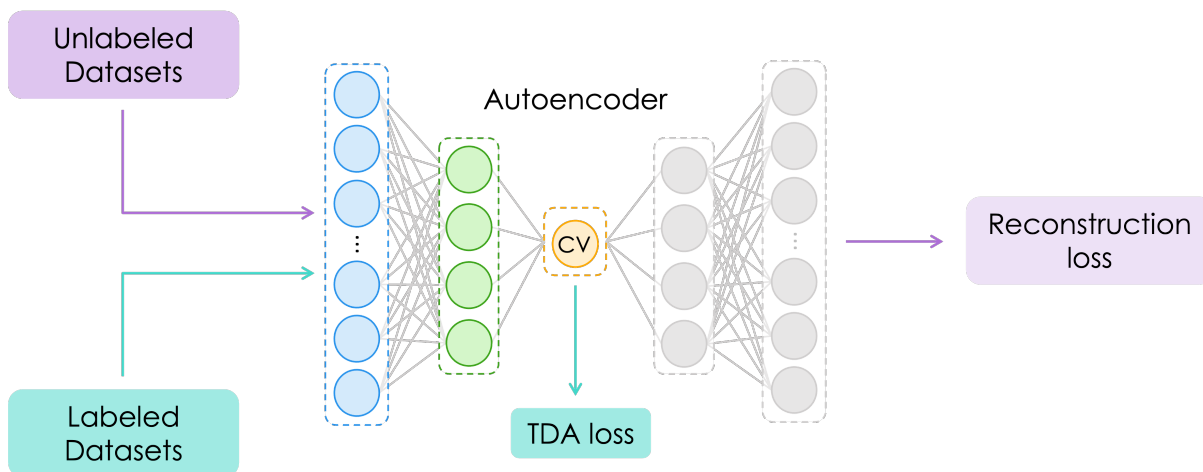


Figure 2.4: **Example of multitask workflow for collective variable (CV) design using semi-supervised learning.** The input features (blue) are derived from multiple datasets, typically, labeled data from classical MD simulations and unlabeled data from biased or reactive trajectories. Data from unlabeled datasets (purple) contributes to the reconstruction loss function, which depends on the decoder’s output. By contrast, data from labeled datasets (cyan) contributes to a supervised loss function, which is the output of the encoder (here, the TDA loss). The resulting CVs (yellow) capture essential dynamical features while remaining grounded in physical interpretability. This iterative scheme allows for progressive refinement of the CVs as new configurations are generated through biased simulations.

Semi-supervised learning represents a hybrid approach that combines both labeled and unlabeled data to improve the learning process, particularly in situations where labeled data are scarce or expensive to obtain. In molecular dynamics (MD), this framework is especially useful when only a limited set of configurations can be confidently assigned to specific states, such as metastable conformations A and B identified through long classical MD simulations, while a larger pool of configurations remains unlabeled. These unlabeled samples often arise from biased or enhanced sampling trajectories that efficiently explore intermediate or transition regions, but lack explicit labels due to their complex or unknown dynamical significance.^{27,172} By incorporating both sources of information, semi-supervised learning allows models to generalize beyond the labeled subset, exploiting the statistical structure and correlations present in the unlabeled data (see right panel in Figure 2.3). This is particularly relevant in biomolecular systems, where transitions between states occur over high-dimensional free energy landscapes, and manually identifying all relevant conformational states are often infeasible. Semi-supervised approaches can thus assist in learning reaction coordinates or collective variables (CVs) that capture essential dynamical features while remaining grounded in physical interpretability. One practical example is the use of neural networks trained to both classify known states (e.g., folded/unfolded, bound/unbound) and to learn a latent low-dimensional representation of the entire dataset. During training, the model is simultaneously optimized to minimize classification error on the labeled subset and to preserve the geometric or kinetic structure of the full data distribution, often through loss functions inspired by unsupervised techniques such as autoencoders or variational inference.¹⁷³ This iterative scheme is often employed in workflows where the model is periodically updated as new configurations are generated through biased simulations, allowing for progressive refinement of the CVs (see Figure 2.4). In this context, semi-supervised learning effectively

bridges the gap between supervised models, limited to predefined labels, and unsupervised models—which may miss critical system-specific features. The resulting CVs are typically more robust and transferable across simulation conditions, and can facilitate both mechanistic interpretation and sampling efficiency in subsequent simulations. Recent applications have demonstrated that semi-supervised strategies can uncover hidden intermediate states, improve the accuracy of transition state identification, and accelerate convergence in enhanced sampling protocols.^{27,172} As such, they represent a promising direction in the design of data-driven CVs, especially for systems where only partial knowledge of the landscape is available.

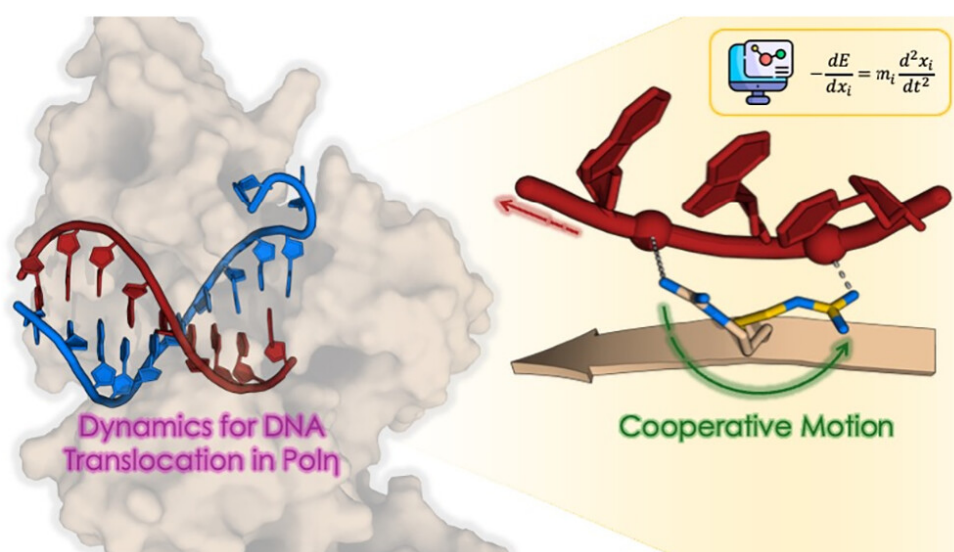
Chapter III – DNA Translocation in wild-type Pol η

This chapter is based on:

Coordinated Residue Motions at the Enzyme–Substrate Interface Promote DNA Translocation in Polymerases, published in *Journal of American Chemical Society* (2025).

Alessia Visigalli, Enrico Trizio, Luigi Bonati, Pietro Vidossich, Michele Parrinello, and Marco De Vivo.

DOI: <https://doi.org/10.1021/jacs.5c05888>.



With the aim of elucidating the mechanism of DNA translocation bypass polymerases, we start by investigating the wild-type human DNA polymerase η (Pol η), developing a framework that combines equilibrium molecular dynamics (MD) simulations with deep-learning-guided enhanced sampling methods. We show that the translocation of the newly elongated double-stranded DNA occurs via a dynamic mechanism where the two DNA strands shift asynchronously along the enzyme. This motion is driven by a set of positively charged residues at the Pol η •DNA interface, which operate in a coordinated way to promote DNA translocation. Our findings provide new insights into the complex yet fundamental dynamic process of DNA translocation in Pols, which is essential for DNA polymerization and thus for the storage and transmission of genetic information in all living organisms.

3.1 Introduction

The DNA translocation in polymerases (Pols) is fundamental to guarantee the preservation of genetic information in all living organisms, allowing DNA replication, transcription, and repair.^{9,11} This process occurs at every catalytic cycle, after nucleotide addition (see Fig. 3.1), and is essential to allow the enzyme to access the next template base for the next nucleotide incorporation (see Figure 1.4).^{92,93,96} That is, the enzyme must translocate the newly elongated double-stranded DNA (dsDNA) by one base pair (bp) along the enzyme, so that the next template base is positioned in the active site of the enzyme.^{7,106} Thus, the translocation process is a key step in the catalytic cycle of Pols, and its understanding is crucial for elucidating the mechanisms of DNA replication and repair. However, the molecular details of this process are poorly understood, and the mechanisms by which Pols translocate DNA remain a topic of active research.^{27,36,37}

Here, we investigate the mechanism of DNA translocation in wild-type human DNA polymerase η

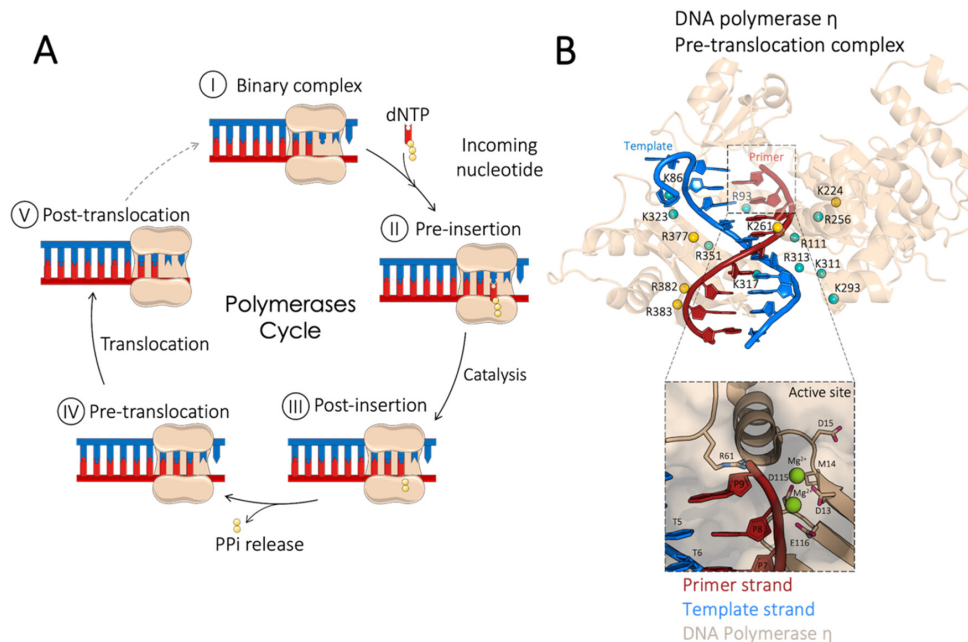


Figure 3.1: (A) **Schematic representation of the Pol η catalytic cycle.** (I) the binary complex Pol η -DNA_n, (II) the ternary complex with the nucleotide located in the active site, (III) the binary complex formed after catalysis, (IV) the binary complex with the active site occupied and without the ions and the PPi, (V) the binary complex after the translocation. (B) **Crystal structure of the DNA polymerase η (Pol η) in the pre-translocation state (PDB ID: 4ECX).** The key positive residues interacting with the DNA are shown in spheres and colored according to the strand they interact with (template: teal; primer: gold). Protein and nucleic acid are represented in cartoon, with different colors for the primer strand (red) and the template strand (blue). Close view of the active site, where the primer strand (red) is in cartoon representation, the active site residues are in sticks, and the magnesium ions are in spheres.

(Pol η), a member of the Y-family of DNA polymerases,^{5,7} which is involved in translesion synthesis (TLS) and plays a crucial role in DNA repair, recently becoming a promising target for cancer therapy.^{13,22,33,89,185} Pol η represents a model system to study the translocation process, as it has been extensively characterized both experimentally and computationally.^{7,33,34,43} Recent time-resolved X-ray structures of Pol η in complex with DNA have been solved, providing a complete picture of the enzyme

in different states along the catalytic cycle, including before and after DNA translocation.^{5,34} These structures have revealed a minimal structural rearrangement of the enzyme within the catalytic cycle, in contrast to replicative polymerases, where the rotation of the fingers subdomain plays a crucial role in catalysis and translocation.^{36,37}

To investigate the mechanism of DNA translocation in Pol η , we develop a framework that combines classical molecular dynamics (MD) simulations with deep-learning-guided enhanced sampling techniques, providing a comprehensive view of the translocation process in the wild-type Pol η . Since structural data inherently lack information on the dynamics of the system, we employed classical MD simulations initiated from experimentally resolved structures, allowing the exploration of the conformational space of the enzyme and its interactions with DNA (see section 2.1). To further characterize the translocation process, we employed the deep-learning-guided enhanced sampling methods (subsection 2.2.1). Our framework provides a detailed understanding of the dynamic mechanism of DNA translocation in Pol η , revealing a previously uncharacterized dynamic process where the two DNA strands shift asynchronously along the enzyme. This asynchronous motion is driven by a set of positively charged residues at the Pol η •DNA interface, which operate in a coordinated way to promote DNA translocation. Notably, the involved residues are functionally relevant and evolutionally conserved across the Y-family of DNA polymerases, suggesting a common mechanism for DNA translocation in repairing Pols.

3.2 Identification of Key Residues Promoting DNA Translocation

To investigate the mechanism of DNA translocation in Pol η , we performed classical MD simulations starting from the X-ray structure of the enzyme in complex with DNA in pre-translocation (PDB ID: 4ECX) and post-translocation (PDB ID: 4ED8) states.³⁴ Despite the overall stability of both models (see Figure A.1 and Figure A.2), our simulations revealed a dynamic rearrangement at the level of the Pol η •DNA interface, influencing the interactions between the enzyme and the DNA substrate. We observed a significant increase in the flexibility of strategically positioned positively charged residues at the Pol η •DNA interface. These fluctuations were particularly pronounced in the pre-translocation state (Figure A.3), where the enzyme is poised to translocate the newly elongated dsDNA. Comparing the two states, we noticed side chain flipping of those residues, allowing them to transiently reach the post-translocation conformation (Figure 3.3). This dynamic rearrangement suggests a spontaneous tendency of the enzyme toward translocation. To further investigate the role of these residues in promoting DNA translocation, we performed three replicas of 2 μ s classical MD simulations starting from the pre-translocation state (PDB ID: 4ECX) and post-translocation state (PDB ID: 4ED8). In both systems, we observed that the DNA maintained its overall structure, with the phosphate backbone and the base pairs remaining intact. The movement of the DNA substrate was limited to small fluctuations around the initial position, with an average root-mean-square deviation (RMSD) ~ 2.5 Å for template and primer strands (see Figure A.4).

3.2.1 Primer•Pol η Interactions

Our molecular dynamics simulations reveal that six positively charged residues at the Pol η •Primer interface (Arg61, Lys224, Lys261, Arg377, Arg382, and Arg383; see Figure 3.2 A left) play a dynamic role in modulating DNA binding during translocation. Among them, Arg382 and Arg383 exhibit pronounced

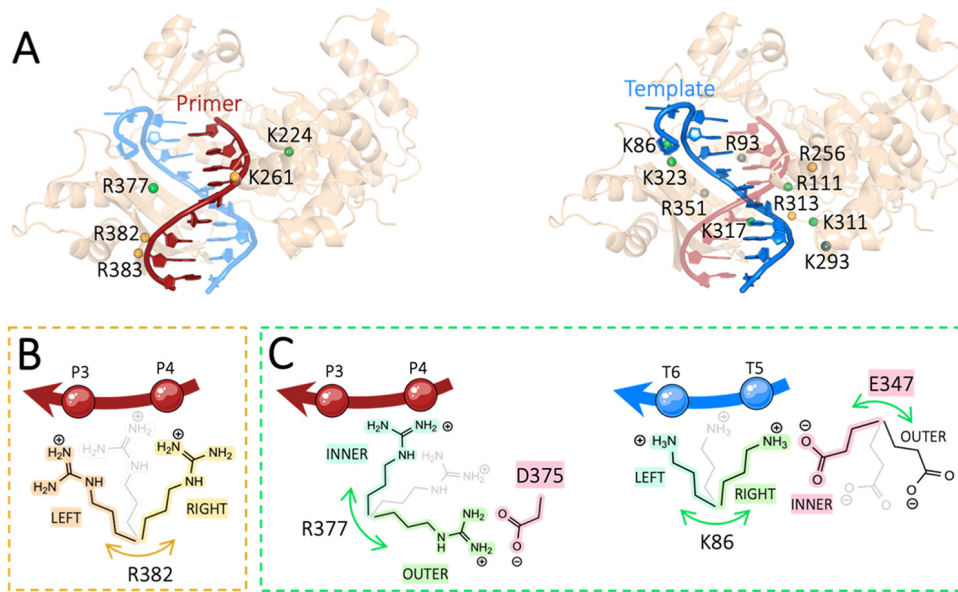


Figure 3.2: (A) **Overview of positive residues at the Pol η •DNA interface.** On the left, the positive residues interacting with the primer strand (red). On the right, the positive residues interacting with the template strand (blue). The positively charged residues are color-coded according to the oscillatory motion explained in panels (B) and (C). (B) Schematic representation of the motion of the R382 side chain, which transiently interacts with two different phosphate atoms of the primer strand. (C) Schematic representation of R377, which changes its initial inner conformation to an outer one, unbinding the primer phosphate P4 in favor of the negative residue D375 (left panel). Another oscillatory motion is schematically represented for K86, which alternatively interacts with the phosphates T6 or T5 and E347 (right panel).

side-chain rearrangements, alternately interacting with different phosphates of the primer strand (Figure 3.2 B). Arg382 shifts between P2, P3, and occasionally the base T9, while Arg383 alternates between P2 and P3. These fluctuations occur despite the absence of primer binding for these residues in the crystal structure, suggesting a role in dynamically guiding strand movement. Lys224 and Arg377 exhibit different behavior, transitioning from DNA contacts to interactions with negatively charged protein residues (e.g., Glu116 and Asp375, respectively), thereby reducing their electrostatic anchorage to the DNA (Figure 3.2 C). Lys261 remains consistently bound to phosphate P5, highlighting its stabilizing function during and after strand movement. Arg61 alternates between two conformations—binding either phosphate P9 (A-conf) or the base T3 (C-conf), with the latter orientation being previously linked to pyrophosphate release after catalysis.⁹⁶ Comparison of early and later simulation stages indicates that several of these primer-contacting residues spontaneously adopt conformations resembling those observed after translocation, supporting their involvement in facilitating strand progression (Figure 3.3).

3.2.2 Template•Pol η Interactions

The template strand engages with a larger network of eleven positively charged residues (Lys86, Arg93, Arg111, Lys293, Lys311, Lys317, Lys323, Arg313, Arg351, Arg371, and Lys376), nearly twice as many as those interacting with the primer strand (Figure 3.2 A right). Many of these residues exhibit side-chain flexibility, enabling transient shifts between phosphate contacts and intra-protein interactions. Lys86 and Arg111, for example, can switch between binding the template backbone and nearby negatively charged

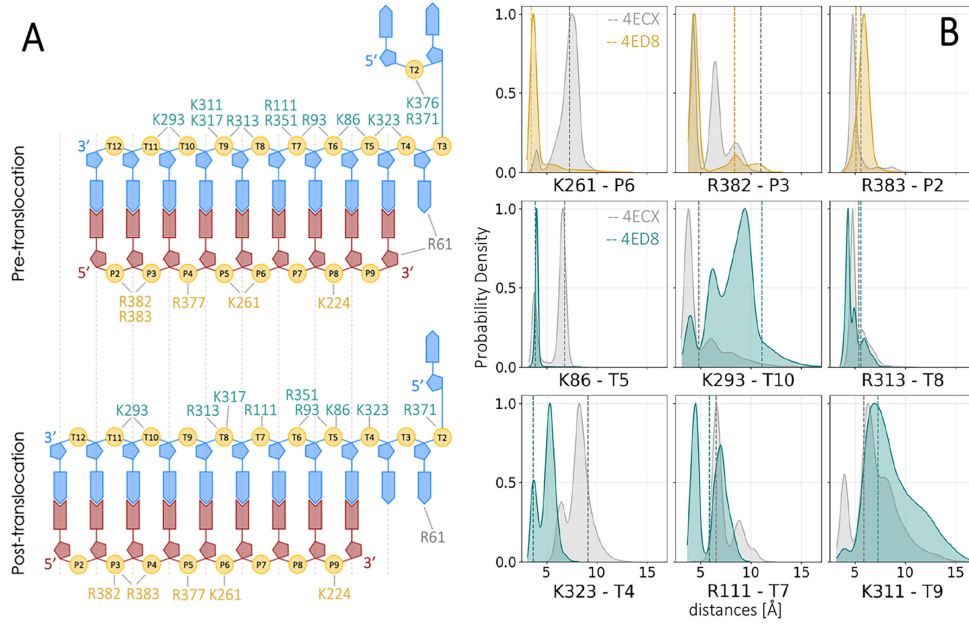


Figure 3.3: (A) **Schematic representation of the Pol η -DNA interactions** in the pre-translocation (upper panel) and post-translocation (lower panel) state. The primer (red) and template (blue) strands are shown, with positive residues labeled in gold (primer-side contacts) and teal (template-side contacts). (B) **Probability distributions of key residue-DNA distances**, highlighting the conformational dynamics in the pre-translocation (gray curves) and post-translocation (colored curves) state. Distances for primer-side interactions (gold) and template-side interactions (teal) are shown, indicating the capability of some residues to reach the post-translocation conformation during the pre-translocation dynamics. The dashed lines represent the values of each distance in the X-ray structures (pre-translocation: PDB ID: 4ECX, gray; post-translocation: PDB ID: 4ED8, colored according to the strand interaction).³⁴

residues (such as Glu347 and Ser96), sometimes through water-mediated bridges (Figure 3.2 C). Arg313 maintains stable interactions with bases T8 and T9, while Lys293 alternates between phosphates T10 and T11 (Figure 3.2 B). Lys311, Lys317, and Lys323, positioned in solvent-exposed regions, frequently break their initial DNA contacts to engage protein residues, such as Asp308 or Tyr39 (Figure 3.2 C). Arg371 and Lys376 transiently approach the template junction, accompanied by a motion of the little finger subdomain that enlarges the gate with the finger subdomain. In contrast, Arg93 and Arg351 form stable, persistent interactions with specific template phosphates throughout the simulations. Similar to the primer strand, many template-contacting residues adopt conformations consistent with a post-translocation arrangement, suggesting that these rearrangements prime the DNA for stepwise movement. However, the full translocation event was not captured within the equilibrium timescales explored, likely due to the limited sampling of thermal fluctuations in conventional MD simulations (section 2.2).

3.3 Characterization of the Asynchronous Stepwise Mechanism.

Despite classical MD simulations providing valuable insights into the interactions between Pol η and the DNA substrate, they inherently lack the temporal resolution needed to fully characterize the translocation process. (section 2.2) To promote such rare event, we thus performed enhanced sampling simulations using the On-the-Fly Probability Enhanced Sampling (OPES) method (subsection 2.2.3). This tool, com-

bined with deep-learning-based collective variables (CVs), allowed us to explore the conformational space of the enzyme and its interactions with DNA, providing a detailed understanding of the translocation process. The first step consists of building a 1-dimensional CV able to classify the system in pre- and

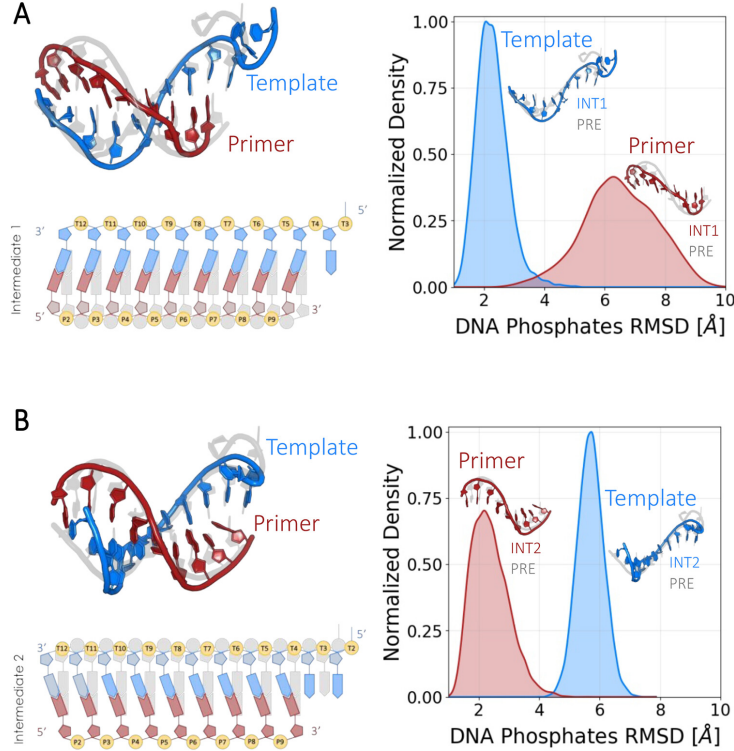


Figure 3.4: Description and classification of the two intermediate states. (A) Intermediate state 1 (INT1). On the left, the structural and the schematic representation of the nucleic acid in INT1. Here, the primer strand (red) is shifted by one base pair with respect to the pre-translocation state (gray). By contrast, the template strand (blue) is in the same position as in the pre-translocation state. On the right, the probability distributions of RMSD values for DNA phosphates in the primer (red) and template (blue) strands, comparing INT1 with the pre-translocation (PRE) state. (B) Intermediate state 2 (INT2). On the left, the structural and schematic representation of the nucleic acid in INT2. Here, the template strand (blue) is shifted by one base pair with respect to the pre-translocation state (gray). By contrast, the primer strand (red) is in the same position as in the pre-translocation state. On the right, the probability distributions of RMSD values for DNA phosphates in the primer (red) and template (blue) strands, comparing INT2 with the pre-translocation (PRE) state.

post-translocation states using Deep-LDA (Figure 2.3.2).¹⁶⁹ As input, we used 69 distances between the DNA substrate and the enzyme, which are changing between pre- and post-translocation state MD, and so they are expected to change during the translocation process. The resulting CV is a continuous function connecting pre- and post-translocation states. We then used this CV to perform multiple OPES simulations, starting from different configurations of pre- and post-translocation states, respectively (Figure A.5). Collecting $\sim 20\mu s$ of independent simulations, comprising more than 80 translocation events, we were able to identify two distinct translocation pathways. Both of them are characterized by a stepwise mechanism, where the two DNA strands shift asynchronously along the enzyme. In the first path, the primer strand translocates first, followed by the template strand (Figure A.6). Here, the system must pass through an intermediate state (INT1) where the primer strand is fully translocated, while the template strand is still in the pre-translocation state. On the other hand, in the second path, the template strand

translocates first, followed by the primer strand. Again, the system must pass through an intermediate state (INT2), opposite with respect to INT1, where the template strand is fully translocated, while the primer strand is still in the pre-translocation state. This asynchronous motion is driven by a set of positively charged residues at the Pol η -DNA interface, which operate in a coordinated way to promote DNA translocation. The involved residues are functionally relevant and evolutionally conserved across the Y-family of DNA polymerases, suggesting a common mechanism for DNA translocation in repairing Pols (Figure 3.7).

3.3.1 Implementing an Optimized Collective Variable for DNA Translocation.

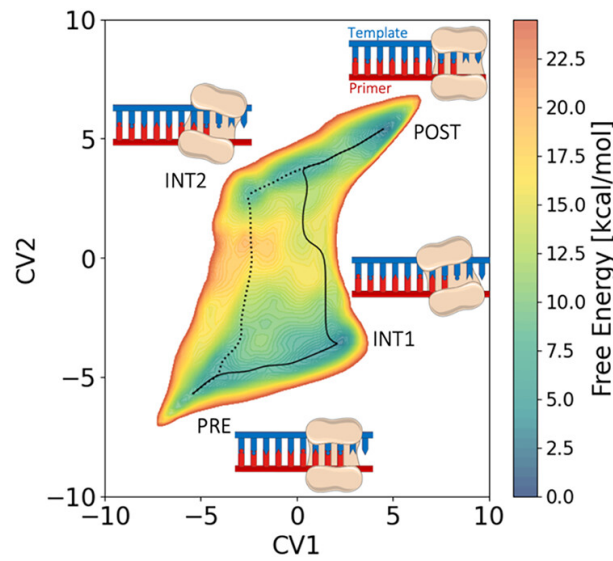


Figure 3.5: **Free energy surface (FES) describing the two translocation paths:** this FES represents two specific translocation pathways where Pol η translocates from the PRE state to the POST state, passing through one of the two intermediate states (INT1 or INT2). The PR state is located in the lower left corner of the FES, which is the starting point of the translocation process, with the DNA in a pre-translocated conformation. The POST state is positioned at the upper right of the FES, representing the final state of translocation, where the DNA has fully shifted to a post-translocation position. The dashed line traces the minimal energy path (pathway 1) between these states, showing how the system passes through energy barriers as it moves from the PR state, through INT1, to the POST state. The dotted line highlights a different translocation pathway (pathway 2) where the polymerase transitions from PRE to POST, passing through INT2.

Having identified the two translocation pathways, we aimed to include these new biological insights in the CV, improving the sampling of the translocation process. First, we ran additional classical MD simulations starting from the INT1 and INT2 states, to collect more data on the system in those intermediates (Figure A.7). After checking the overall system stability, we employed a multitask framework to train a 2-dimensional CV (Figure 2.3.2). This approach is motivated by the intention to include in the NN input the data coming from both classical MD simulations in the four states (pre- and post-translocation, INT1 and INT2), as well as data collected in the previous OPES simulations. The resulting CV spatially separates the two translocation pathways, allowing us to distinguish between the two asynchronous steps. We then used this new CV to perform additional OPES simulations, starting from different configura-

tions of pre- and post-translocation states, respectively. Collecting new OPES simulations using the new CV, we were able to sample the translocation process in both pathways, confirming the asynchronous stepwise mechanism. Our simulation revealed the propensity of the system to translocate the primer strand first, followed by the template strand. This result was then confirmed by the free energy surface (FES) calculated from the OPES simulations, which showed a lower free energy barrier for the primer strand translocation compared to the template strand. The FES also revealed the presence of two distinct minima corresponding to the INT1 and INT2 states, further supporting the asynchronous stepwise mechanism.

3.4 Discussion

Recent time-resolved X-ray crystallography studies have, for the first time, captured complete DNA synthesis events in human DNA polymerase η (Pol η), revealing structures before and after DNA translocation.³⁴ Intriguingly, these structures displayed no major conformational rearrangements in the enzyme between the two states; the primary structural difference was the position of the DNA substrate. This observation left the dynamic mechanism of DNA translocation in Pol enzymes unresolved. In this work, we combined microsecond-long molecular dynamics (MD) simulations with deep-learning-guided enhanced sampling to investigate the wild-type Pol η •DNA complex in its initial (pre-translocation) and final (post-translocation) states. Our equilibrium MD simulations ($\sim 2 \mu\text{s}$ per replica) revealed pronounced side-chain motions of specific positively charged residues at the Pol η •DNA interface. These rearrangements suggest a mechanism in which such residues transiently grasp the DNA double strand and facilitate its sliding through the binding cleft. Grouping the residues according to their primary interaction partner revealed that four of the six residues anchoring the primer strand (Arg61, Lys261, Arg382, and Arg383) and six of the eleven residues anchoring the template strand (Lys86, Arg111, Lys293, Arg313, Lys323, and Arg371) adopted conformations in the pre-translocation simulations that closely resembled those in the post-translocation structure (Figure 3.3). This behavior supports the hypothesis that their coordinated “windshield wiper”-like motion assists the stepwise progression of the DNA strands during translocation. To investigate this mechanism in detail, we performed enhanced sampling simulations using machine-learning-based collective variables, allowing us to collect 81 transitions between pre- and post-translocation states. The resulting free-energy landscapes revealed two mutually exclusive, competitive pathways, both following an asynchronous, stepwise mechanism (Figure 3.6). Each pathway featured a distinct intermediate, INT1 or INT2, differing in the order of strand movement (Figure 3.4). In INT1, the primer strand shifted first while the template strand remained in its pre-translocation position; in INT2, the order was reversed. Energetically, translocating the primer strand first required $\sim 7.5 \text{ kcal mol}^{-1}$, whereas moving the template strand first required $\sim 17 \text{ kcal mol}^{-1}$ (Figure 3.5). This difference reflects the larger number of stabilizing contacts anchoring the template strand, consistent with its role as the positional reference for nucleotide incorporation. While our simulations do not explicitly capture pyrophosphate release, prior computational work suggests that chemical and physical steps in Pol catalysis are separable, with Mg^{2+} ions dissociating before translocation.³⁶ We observed transient binding of divalent cations near the catalytic site, but their possible influence on translocation remains speculative. Analysis of strand-specific RMSD changes along the pathways supports the proposed mechanism. In

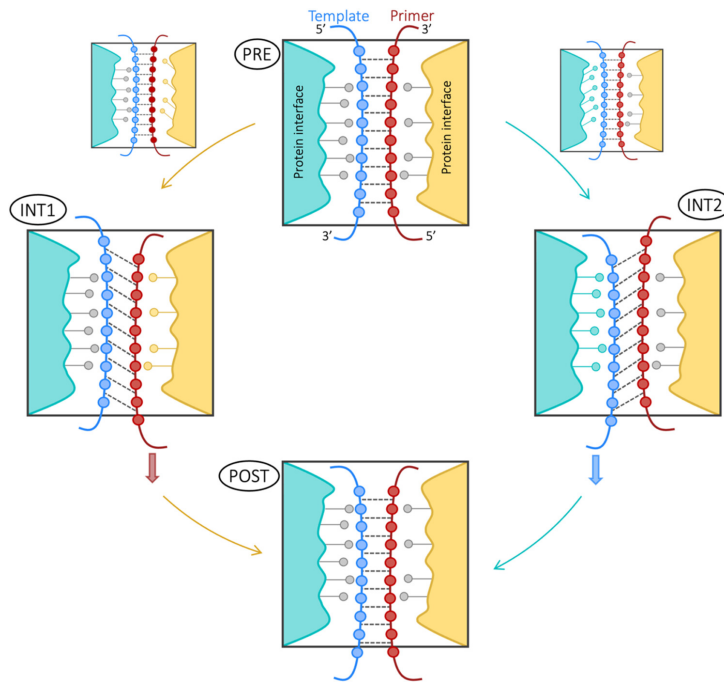


Figure 3.6: **Schematic representation of the asynchronous DNA translocation mechanism in Pol η .** The transition from the pre-translocation (PRE) to the post-translocation (POST) state occurs via two mutually exclusive pathways, i.e., path1 on the left and path2 on the right. Each path is characterized by an intermediate state (INT1 or INT2). In INT1, the primer strand (red) is fully translocated, adopting the interactions typical of the POST state, while the template strand (blue) retains the PRE state configuration. Conversely, in INT2, the template strand is fully translocated, while the primer strand maintains its PRE state interactions. The final step leads to the fully translocated POST state, where both strands adopt their final positions. Arrows indicate the direction of strand movement.

the PRE $\rightarrow$ INT1 transition, the primer strand exhibited a large RMSD increase, while the template strand remained stable. The INT1 $\rightarrow$ POST step showed the opposite pattern, with template movement accompanied by stabilization of the primer. Structural analysis revealed that the template junction is stabilized through stacking interactions between the fingers and little-finger subdomains, in particular by aromatic residues Trp42 and Trp64, which appear to act as a gating mechanism for substrate accommodation (Figure A.8). Conservation analysis across Y-family polymerases shows that, although the little-finger subdomain is the least conserved in sequence, most positively charged residues anchoring the DNA are structurally conserved, underscoring their likely functional importance (Figure 3.7). Notably, the little-finger subdomain is absent in polymerases employing alternative translocation mechanisms, further supporting its specialized role in this stepwise process.³⁹ Finally, comparison with other polymerases suggests that this asynchronous, strand-specific mechanism may be common among Y-family members, such as Pol ι and Pol κ , which also present a high number of interface positive residues. In contrast, A-family polymerases often rely on large-scale motions of structural elements like the O-helix to trigger translocation, a feature absent in Pol η .^{36,37} Similarly, viral RNA-dependent RNA polymerases (RdRPs) display minimal global rearrangements and have been crystallographically captured in intermediates analogous to INT1, where only the growing strand has shifted.^{90,186,187} Together, these results support a conserved functional role for positively charged interface residues and specialized subdomain

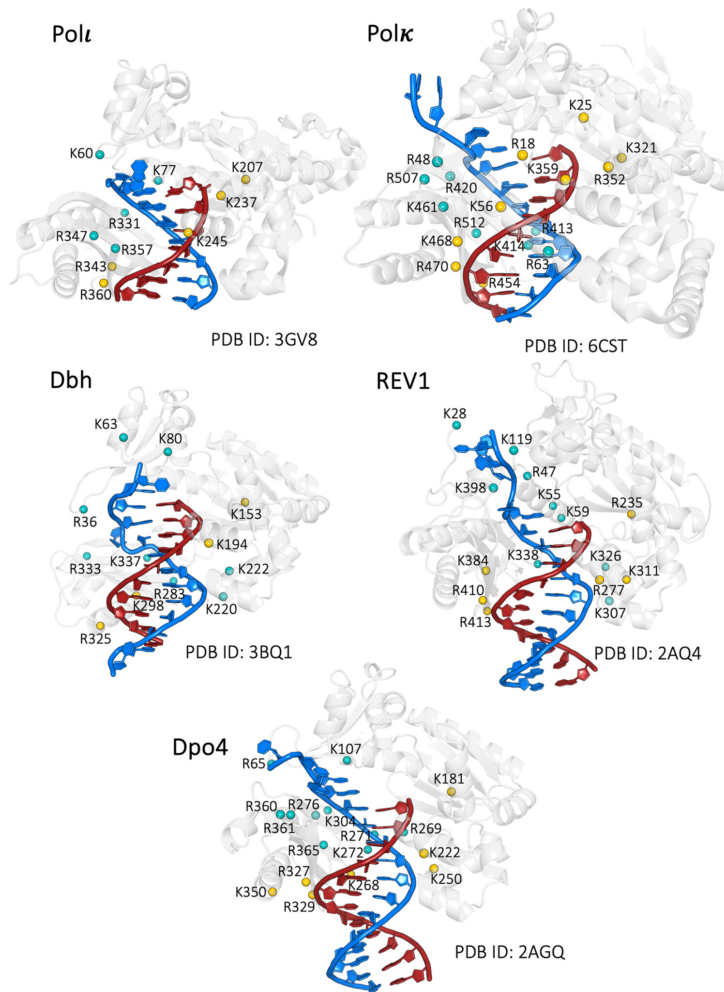


Figure 3.7: **Structural comparison of positively charged residues interacting with DNA in Y-family polymerases.** Crystal structures of Pol ι (PDB ID:3GV8), Pol κ (PDB ID: 6CST), Dbh (PDB ID: 3BQ1), Rev1 (PDB ID: 2AQ4), and Dpo4 (PDB ID: 2AGQ) highlighting positively charged residues (Arg and Lys) at the protein-DNA interface. DNA is shown in blue (template strand) and red (primer strand), while positively charged residues are represented as spheres (teal if they interact with the template strand; gold if they interact with the primer strand) and labeled. The structural alignment reveals conserved electrostatic interactions stabilizing the DNA substrate across different polymerases, despite the sequence variability in the little-finger subdomain.

architectures in driving DNA translocation in Y-family polymerases through an asynchronous, stepwise mechanism.

3.5 Conclusions

Our results reveal a previously unrecognized translocation mechanism in DNA polymerase η (Pol η), driven by the coordinated movement of strategically positioned positively charged residues at the Pol•DNA interface. This coordinated motion enables these residues to grasp and guide the DNA substrate through the polymerase binding site. Extensive molecular simulations and free-energy calculations, supported by a broad range of experimental evidence, point to an asynchronous translocation process involving two competing pathways. The preferred route is one in which the primer strand advances first, while the

template strand, more strongly anchored to the enzyme by several positive residues, follows. Structural and sequence conservation analyses suggest that this dynamic, stepwise mechanism is likely shared among Y-family polymerases. Overall, these findings propose a new model for DNA translocation in polymerases, offering experimentally testable predictions to further illuminate this fundamental and complex process in DNA synthesis.

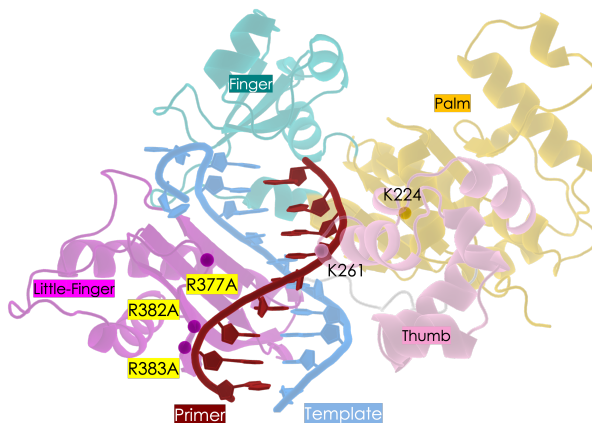
Chapter IV – Arginine Mutations: Functional Role in Translocation

This chapter is based on:

Hybrid Workflow to Understanding Translocation Mutant Systems of Pol η

Alessia Visigalli, Elisa Romeo, Natasha Margaroli, Angela Andonaia, Pietro Vidossich, Stefania Girotto and Marco De Vivo.

Manuscript in preparation.



DNA polymerase η (Pol η) plays a critical role in translesion synthesis, allowing replication past DNA lesions. While numerous studies have characterized the catalytic mechanism of Pol η , the specific molecular determinants of DNA translocation remain poorly understood, particularly in mutant variants. In this chapter, we investigate the role of three arginine residues (Arg377, Arg382, and Arg383) in the little-finger subdomain, which are hypothesized to stabilize the DNA substrate during translocation. Using a hybrid approach combining site-directed mutagenesis, enzymatic assays, and all-atom molecular dynamics simulations, we evaluated the effects of single and double alanine substitutions on polymerase activity, processivity, and DNA-protein interactions. Our results show that Arg382 and Arg383 are critical for maintaining the geometry and electrostatic network that anchors the DNA primer, while Arg377 plays a more minor stabilizing role. The double mutant R382A/R383A exhibits severe destabilization of the DNA substrate and impaired post-translocation conformation, consistent with experimental reductions in processivity. These findings provide a detailed mechanistic understanding of how specific residues contribute to Pol η translocation, offering insights into the cooperative nature of arginine-mediated DNA

stabilization and informing potential strategies for engineering polymerases with tailored processivity.

4.1 Introduction

In this chapter, we focus on three arginine residues (Arg377, Arg382, and Arg383) located in the little-finger subdomain of Pol η . These residues are hypothesized to play a crucial role in stabilizing the DNA substrate during translocation in the wild-type enzyme.²⁷ To investigate their specific contribution, we generated single and double alanine mutants (R377A, R382A, R383A, and R382A/R383A) and assessed their impact on polymerase activity and processivity using experimental assays. Complementary molecular dynamics simulations were performed to elucidate the structural and energetic consequences of these mutations on the translocation mechanism. By integrating experimental and computational approaches, we aim to provide a comprehensive understanding of how these arginine residues influence Pol η function during DNA synthesis.

Although numerous studies on polymerases exist in the literature, often relying on standardized and widely recognized pipelines, quantifying the specific contribution of translocation to overall polymerase activity remains challenging.^{27, 36, 37, 53} Currently, no direct experimental method can separate the energetic effects of translocation from those of DNA binding, and as a result, experimental data on Pol η translocation are limited, particularly for mutant systems.^{53, 188} Identifying mutations capable of modulating translocation could provide valuable opportunities for engineering polymerases for synthetic biology and biotechnological applications.⁷ To address this gap, we applied our computational pipeline to a series of single-point mutations in the Pol η •DNA complex, focusing on positively charged residues at the primer interface that are likely to influence DNA positioning during translocation.²⁷ In parallel, our collaborators at the Italian Institute of Technology (IIT) developed an experimental workflow to measure processivity rates for the wild-type enzyme and four mutant variants. Since processivity reflects contributions from both DNA binding and translocation, computational simulations are essential to disentangle these contributions at the atomistic level. By comparing the results of the mutant systems with the wild-type, we gained new insights into the molecular determinants governing Pol η translocation and identified key residues that play a cooperative role in stabilizing the DNA substrate.

4.2 Experimental Validation of Arginine Mutant Activity and Processivity

To quantify the contribution of specific residues to the translocation of Pol η along the DNA, we have developed a hybrid experimental-computational strategy. The experiments were performed by our collaborators at the Structural Biology Facility at the IIT in Genova (SBF lab; Dr. Stefania Girotto). The constructs of POLH containing the selected mutations were generated through site-directed mutagenesis. Protein expression and purification were performed using the same procedure established for the wild-type enzyme, which involves a single affinity chromatography step to obtain highly pure and homogeneous samples. Among the tested mutants, R377A displayed enzymatic activity comparable to that of the wild-type POLH, indicating that this substitution does not markedly affect the catalytic properties of the enzyme. In contrast, the R382A, R383A, and R382A/R383A mutants exhibited a substantial

decrease in activity. Notably, this difference became even more evident at lower enzyme concentrations, suggesting that the reduced activity may arise from a lower binding affinity or decreased processivity.

Functional evaluation was carried out using two complementary assays. The primer extension assay provides a measure of overall polymerase activity, combining the effects of DNA binding, nucleotide incorporation, and translocation (Figure 4.1 left).⁵³ In contrast, the processivity assay, which employs a DNA trap to prevent re-binding to new DNA molecules, estimates the number of nucleotides incorporated during a single binding event, thereby reducing the contribution of catalysis to the measurement (Figure 4.1 right).⁵⁶ Experimental results indicate that most mutations, with the notable exception of R377A, significantly decrease both overall activity and processivity, with a ranking of:

$$WT \approx R377A > R383A \geq R382A \gg R382A/R383A$$

These findings suggest that the affected arginine residues play an important role in maintaining efficient DNA translocation along the polymerase. Although these assays provide valuable functional information,

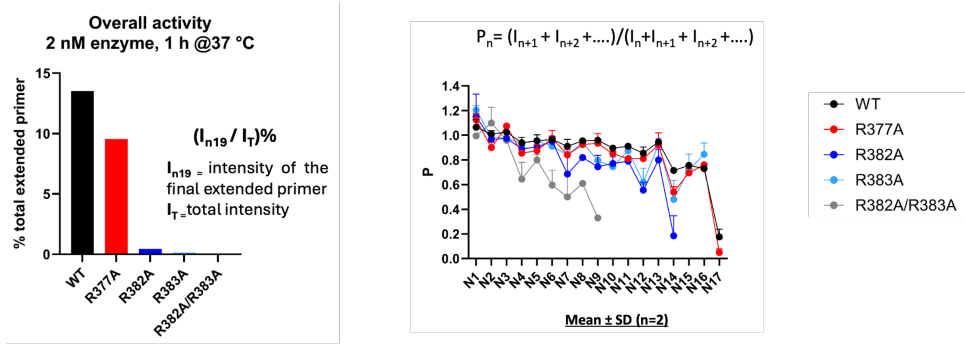


Figure 4.1: **Preliminary experimental results of Pol η mutated systems** (Left) Experimental data from the primer extension assay for the wild-type and mutant Pol η enzymes. The primer extension assay measures the overall polymerase activity, including DNA binding, nucleotide incorporation, and translocation. (Right) Experimental data from the processivity assay for the wild-type and mutant Pol η enzymes. Processivity is calculated as mentioned in Equation 1.3.⁵³

they do not allow a direct quantification of the energetic contribution of translocation to the observed effects. To address this, we complement the experimental results with all-atom molecular dynamics simulations and enhanced sampling techniques. These simulations are used to estimate free energy profiles associated specifically with DNA translocation along the enzyme, providing a way to quantify the energetic barriers that each mutation may introduce. By comparing simulated translocation energies with experimental processivity data, we aim to discriminate between effects on translocation, DNA binding, and catalysis.⁵⁶ The combination of these approaches, still under development, offers a framework to systematically identify residues critical for efficient Pol η translocation and to guide future experimental and computational investigations.

4.3 Computational insights into Arginine Mutations in DNA translocation

To investigate the structural and dynamic effects of the Arg377, Arg382, and Arg383 mutations on Pol η , we performed molecular dynamics simulations starting from pre- and post-translocation states. The analysis focuses on the local rearrangements of the little-finger subdomain, the stability of the Pol η •DNA interface, and the positioning of the DNA duplex within the active site cleft. This approach allows us to examine how the removal of positively charged residues affects the electrostatic network coordinating the primer strand and to identify potential consequences for DNA translocation.

4.3.1 Arg377Ala: Structural Compensation and DNA Misalignment

Arg377 is located on a loop in the little-finger subdomain and interacts with Asp375 on the same loop, as well as with the phosphate backbone of the primer strand (Figure 4.2). In particular, we found that the side chain flipping of Arg377 allows it to interact alternately with Asp375 and P3 in the wild-type system (subsection 3.2.1).²⁷ Despite the poor sequence conservation of the little-finger subdomain, the loop containing Arg377 is structurally conserved among Y-family polymerases.

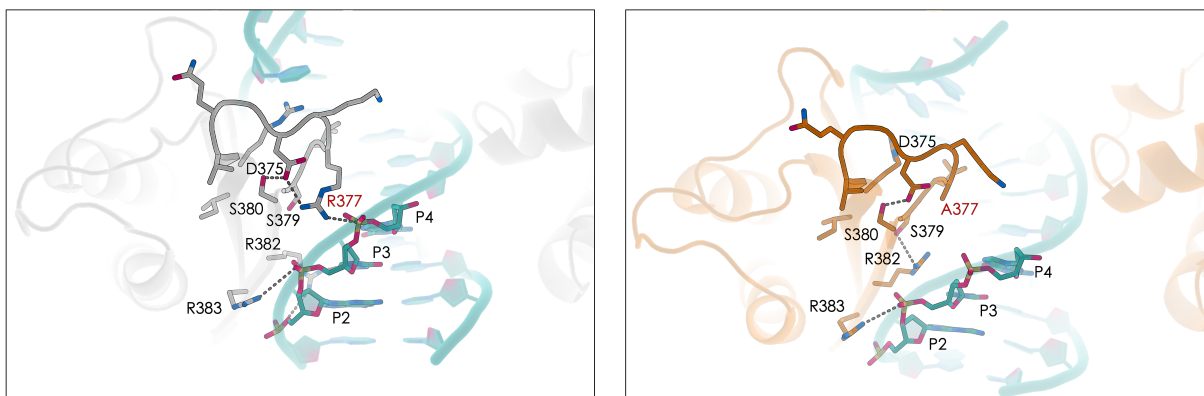


Figure 4.2: **Structural impact of Arg377Ala mutation in Pol η .** (Left) Wild-type Pol η structure highlighting Arg377 and its interactions with Asp375 and the DNA primer base P3. The loop containing the mutation is shown in cartoon and sticks. (Right) Structural representation of the Arg377Ala mutant, showing the loss of interactions with Asp375 and the DNA primer strand.

Arg377Ala was the only mutant that did not show a significant decrease in enzymatic activity compared to the wild-type enzyme, suggesting that this residue plays a minor role in stabilizing the DNA substrate during translocation. To further investigate this, we performed three replicas of 2 μ s classical MD simulations starting from the pre- and post-translocation states of the Arg377Ala mutant (Appendix B).

Our simulations revealed that the overall structure of the Pol η •DNA complex remained stable, with RMSD values comparable to those observed in the wild-type system (Figure B.2). The DNA substrate maintained its integrity and the active site organization was not distorted.

Notably, most trajectories sampled a conformation in which the DNA template was slightly shifted relative to the wild-type reference (RMSD 5.5 ± 0.3 Å), despite Arg377Ala maintaining contacts with the primer strand. This behavior can be rationalized by examining the electrostatic network at the

Pol η •DNA interface. In the wild-type system, five positively charged residues (Lys224, Lys261, Arg377, Arg382, and Arg383) interact with the primer strand (chapter 3).²⁷ Arg377 lies in the same subdomain as Arg382 and Arg383. The loss of Arg377 may be partially compensated by these two arginines, which preserve sufficient electrostatic interactions with the DNA substrate.

However, the absence of Arg377 allows Arg382 to interact with Ser379, located on the same loop. This alternative interaction network prevents Arg382 from stabilizing the primer strand, leading to a slight shift of the DNA substrate (Figure 4.2, right). Consequently, the DNA is not perfectly aligned within the active site cleft, exhibiting a small rotation along its Z-axis. This misalignment likely results from the loss of stabilizing contacts between Arg377 and Arg382 and the primer strand.

To compensate for this loss, the little-finger subdomain moves slightly toward the DNA to better embrace the double helix. Although this configuration seems to prevent DNA unbinding, it weakens the interactions between the finger and little-finger subdomains. This effect is particularly evident in the pre-translocation state, where the little-finger moves away from the finger subdomain, reducing enzyme-template junction contacts.

Overall, these effects promote a slight shift of the template strand. Indeed, Lys323, Arg93, and Arg351 increase their interactions with the template strand, forming an alternative network that stabilizes the DNA in the absence of Arg377 and may facilitate template translocation. However, this rearrangement also leads to a misalignment of the DNA within the active site cleft, particularly involving the primer bases near the catalytic site (P5-P9), whereas the bases interacting with the little-finger subdomain maintain their wild-type positions. This misalignment is evident in the distances at the Pol η •DNA interface involving the primer bases from P5 to P9 (Figure B.3).

In contrast, the post-translocation state simulations of Arg377Ala showed a most-populated cluster resembling the wild-type conformation (primer RMSD: 3 ± 0.8 Å; template RMSD: 3.6 ± 0.7 Å), with the DNA substrate properly aligned within the active site cleft and the little-finger subdomain maintaining its interactions with the finger subdomain. Interestingly, the Pol η •DNA interface distances involving the finger subdomain were particularly unstable, frequently fluctuating among clusters, whereas the distances involving the little-finger subdomain were stable and comparable to the wild-type post-translocation state.

This behavior suggests that the little-finger subdomain plays a critical role in stabilizing the DNA substrate in the post-translocation state, while the finger subdomain remains more flexible and adaptable. Additionally, our simulations revealed a hybrid conformation comparable to the intermediate 2 observed in the wild-type system (Figure 3.4).²⁷ In this intermediate, the primer strand is post-translocated while the template strand remains in a pre-translocation position. The interaction network between the thumb subdomain and the primer strand in this cluster resembles that of the pre-translocation state.

4.3.2 Arg382Ala and Arg383Ala: Differential Roles in DNA Stabilization and Translocation

Arg382 and Arg383 are located on the same loop of the little-finger subdomain, adjacent to the α -helix H. Together with Arg377, these residues form a positively charged cluster that interacts with the phosphate backbone of the primer strand near bases P2 and P3, contributing to the stabilization of the DNA duplex during translocation. Mutating these residues to alanine removes key electrostatic contacts in a region that plays a pivotal role in maintaining the correct alignment of the DNA within the active site cleft.

In both Arg382Ala and Arg383Ala mutants, the little-finger subdomain exhibits a rotation opposite to that observed in Arg377Ala: the α -helix H shifts upward, while the loop carrying the mutation moves downward toward the DNA substrate. This rearrangement likely reflects a compensatory motion to restore the lost electrostatic stabilization near the primer phosphates.

In the pre-translocation state, both mutants display similar structural responses. The little-finger subdomain moves closer to the DNA, leading to a tightening of the local binding pocket and a partial reduction in the distance between the primer and the protein. The primer remains overall stable, but the duplex exhibits a subtle tilting relative to the helical (Z) axis, consistent with a mild misalignment of both strands within the active site cleft.

In the post-translocation state, however, clear differences emerge between the two systems. The Arg382Ala mutant exhibits a markedly unstable behavior compared to the wild-type enzyme, with the primer strand only partially translocated and the template strand remaining in a pre-translocation-like position. The resulting hybrid arrangement resembles the intermediate 2 of the wild-type system. The DNA duplex is tilted with respect to the helical axis, such that the region of the primer near the active site is shifted backward (pre-like), while the distal portion near the little-finger subdomain maintains a post-translocation conformation. This misalignment leads to a decoupling between the active site and the little-finger subdomain, although the latter preserves stable, wild-type-like contacts with the DNA (Figure B.6).

By contrast, the Arg383Ala mutant behaves more similarly to the wild-type enzyme (RMSD < 4.5 Å for both strands). The overall conformation of the DNA substrate remains properly aligned within the active site, and no significant structural rearrangements of the protein are observed. The only consistent difference is a slight decrease in the distance between the C α atom of Ala383 and the phosphate group of primer base P3, comparable to that seen in the wild-type pre-translocation state. This minor difference is due to the loss of the Arg383 side chain, which in the wild-type system interacts with the primer phosphate.

Taken together, these results suggest that Arg382 plays a more critical role than Arg383 in coupling the movement of the little-finger subdomain to the translocation of the DNA duplex. While Arg383Ala preserves near-native behavior, the Arg382Ala mutation disrupts the delicate balance between DNA alignment and subdomain coordination, resulting in a tilted, hybrid configuration of the duplex and reduced translocation efficiency.

4.3.3 Arg382Ala/Arg383Ala double mutant promotes DNA unbinding

To further investigate the cooperative role of Arg382 and Arg383 in stabilizing the DNA substrate, we performed additional simulations on the double mutant Arg382Ala/Arg383Ala. This system provides insight into the cumulative effect of removing both positive charges from the little-finger subdomain, which are positioned in close proximity to the phosphate groups of the primer strand near bases P2 and P3.

In this mutant, the little-finger subdomain exhibits a pronounced opening movement relative to the rest of the polymerase (Figure B.7). The upward shift of the α -helix H and the simultaneous downward displacement of the loop carrying the mutations disrupt the hydrogen-bond and electrostatic network that normally anchors the DNA within the enzyme cleft. As a consequence, the DNA duplex loses stable

contacts with the little-finger subdomain and gradually detaches from the protein surface.

During the simulations, the distances between the primer phosphates and the little-finger residues increase significantly, while the interactions involving the thumb and finger subdomains become more transient and uncoordinated. These observations indicate a loss of the tight coupling between the enzyme and the DNA, consistent with an early-stage unbinding event. The overall picture suggests that the presence of at least one of the two positively charged residues, Arg382 or Arg383, is required to preserve the integrity of the Pol η •DNA interface and to prevent premature DNA release.

Taken together, the results from all mutants reveal a cooperative mechanism in which the cluster of arginine residues (Arg377, Arg382, and Arg383) ensures both the alignment and the retention of the DNA substrate during the translocation cycle. While individual mutations can be partially compensated by local rearrangements of the little-finger subdomain, the simultaneous loss of Arg382 and Arg383 disrupts the electrostatic balance, leading to an open conformation of the enzyme and loss of substrate binding.

Collectively, the computational analysis highlights a cooperative role of Arg377, Arg382, and Arg383 in stabilizing the primer strand during the translocation process. While Arg377Ala shows a nearly wild-type behavior, the loss of Arg382, and even more so of both Arg382 and Arg383, compromises the electrostatic network at the DNA interface, leading to partial misalignment and destabilization of the complex. These findings complement the experimental observations, providing a mechanistic explanation for the decreased processivity and activity observed in Arg382Ala and Arg383Ala mutants.

4.4 Discussion

The mutational analysis of Arg377, Arg382, and Arg383 provides new insight into the structural and dynamic role of the little-finger (LF) subdomain in controlling DNA translocation in Pol η . Although these residues are spatially close and belong to the same positively charged cluster, their impact on the Pol η •DNA complex varies markedly, reflecting distinct contributions to DNA stabilization and enzyme conformational coupling.

Among the three single mutants, Arg377Ala exhibited the mildest effect, in line with experimental data showing a minimal reduction in enzymatic activity. Despite the loss of a positively charged side chain, the LF loop containing Ala377 maintained stable interactions with the DNA, and the enzyme preserved its overall conformation. The local rearrangement observed, involving the formation of an alternative electrostatic contact between Arg382 and Ser379, suggests that Arg377 acts as a modulator rather than a key stabilizer of the primer strand. Its mutation can be compensated by neighboring residues, particularly Arg382 and Arg383, which sustain the electrostatic anchoring of the DNA duplex.

In contrast, the Arg382Ala and Arg383Ala mutants revealed a more pronounced reorganization of the LF subdomain, characterized by opposite rotational motions relative to the Arg377Ala case. Both mutants exhibited a tightening of the LF•DNA interface in the pre-translocation state, likely reflecting a compensatory response to the loss of charge near the primer phosphates. However, their post-translocation conformations diverged significantly. Arg382Ala disrupted the proper alignment of the primer and template strands, leading to a tilted, hybrid configuration reminiscent of the wild-type intermediate 2. This indicates that Arg382 plays a central role in coupling the LF dynamics with the translocation of the DNA duplex. Arg383Ala, on the other hand, preserved near-native geometry and

subdomain coordination, supporting the view that Arg383 provides local electrostatic stabilization but is not essential for large-scale subdomain coupling.

The double mutant Arg382Ala/Arg383Ala clearly demonstrated the cooperative nature of these two residues. Removing both charges caused a loss of structural integrity at the Pol η •DNA interface, with the LF subdomain rotating away from the DNA and the complex adopting an open, partially unbound conformation. This structural destabilization suggests that at least one of the two arginines is necessary to preserve the compact architecture required for efficient substrate retention and translocation.

However, since these results are based on equilibrium molecular dynamics simulations, the direct effect of each mutation on the translocation process cannot yet be fully assessed. While classical MD captures local rearrangements and binding stability, it does not explicitly describe the energy barriers or transition pathways associated with strand movement along the enzyme. Therefore, the molecular rationale underlying the difference in catalytic activity between Arg377Ala and the other mutants remains only partially resolved. It is plausible that the mutations affect the energetics or coordination of the translocation step rather than the static stability of the complex, a hypothesis that will be further explored through enhanced sampling simulations.

Overall, these findings support a model in which Arg377, Arg382, and Arg383 collectively form a dynamic electrostatic network that stabilizes the primer strand and ensures the mechanical coupling between the LF subdomain and the translocation machinery. While single mutations can be locally compensated, the simultaneous disruption of this network impairs DNA binding and may hinder processive translocation. These results provide a structural framework consistent with experimental trends and highlight the importance of the LF subdomain as a flexible, yet functionally critical, element of Pol η translocation mechanism.

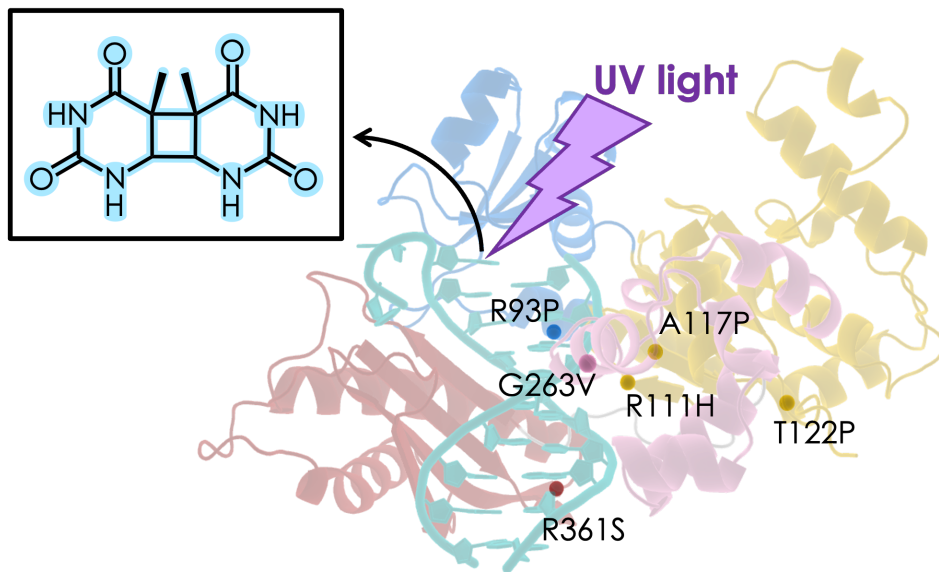
Chapter V – Impact of Pathological Mutation on Pol η Translocation

This chapter is based on:

Decoding the Impact of XP-V Associated Mutations in DNA Polymerase η through Atomistic Simulations.

Alessia Visigalli, Paolo Carloni, and Marco De Vivo.

Manuscript in preparation.



This work aims to elucidate the impact of pathological mutation on DNA polymerase η (Pol η) activity. Based on our previous study regarding DNA translocation in wild-type system (chapter 3), we characterized the pathological mutated form of Pol η , using classical MD techniques. Comparing structural and dynamical features of six pathogenic variants (Ala117Pro, Thr122Pro, Gly263Val, Arg93Pro, Arg111His, and Arg361Ser) with the wild-type enzyme in both pre- and post-translocation states, we highlight the impact of these mutations on the enzyme's function. The findings highlight the intricate balance between enzyme flexibility and stability, providing insights into the molecular basis of catalytic or translocation defects in Pol η . This work contributes to our understanding of DNA repair mechanisms and the role of Pol η in maintaining genomic integrity, with implications for cancer research and therapeutic strategies.

5.1 Introduction

Lesion bypass by non-replicative DNA polymerases (Pols) is a crucial mechanism that allows cells to tolerate DNA damage and maintain genomic stability. During translesion synthesis (TLS), specialized enzymes insert nucleotides opposite damaged bases, enabling replication to continue despite the presence of lesions. This process prevents replication fork stalling and ensures faithful transmission of genetic information. Defects in TLS can lead to severe consequences, including genomic instability, cancer predisposition, and developmental disorders. In eukaryotes, most TLS polymerases belong to the Y family and are characterized by low fidelity and processivity. Among them, DNA polymerase η (Pol η) plays a key role in bypassing UV-induced lesions, such as cyclobutane pyrimidine dimers (CPDs) and cisplatin adducts.^{189,190} Deficiency in Pol η activity causes Xeroderma Pigmentosum Variant (XP-V), a rare genetic disorder marked by hypersensitivity to UV light and a high incidence of skin cancer.^{13,89,191}

Recent crystallographic and computational studies have elucidated the molecular basis of Pol η lesion-bypass activity, highlighting the importance of specific residues in nucleotide binding, catalysis, and translocation along the DNA substrate.^{34,36,37} However, most of these investigations focused on the wild-type enzyme, leaving a gap in our understanding of how pathogenic variants alter Pol η activity and at which step of the catalytic cycle they act. Experimental data show that Pol η efficiently bypasses CPDs located at the first or second position of the DNA template but is much less efficient at lesions in the third or fourth position.³³ This position-dependent behavior reflects changes in processivity and fidelity: Pol η can switch between low and high processivity modes depending on the lesion's location, likely to facilitate the transition back to a replicative polymerase after bypass.¹²⁵ Despite reduced catalytic efficiency, the enzyme maintains increased processivity and fidelity in the presence of CPDs, suggesting that tighter protein-lesion interactions contribute to accurate bypass.³³ Pol η also tolerates other lesions, such as cisplatin adducts, though with lower efficiency, indicating a preference for specific lesion types. Interestingly, it can function as a reverse transcriptase, efficiently accommodating RNA/DNA hybrid substrates. This promiscuity stems from its spacious active site, which allows bulky substrates and mismatched base pairs to be processed without compromising activity.^{33,76} Pol η thus represents a model system for understanding TLS mechanisms in Y-family polymerases.⁷⁶ X-ray structures capturing multiple catalytic states have shed light on its conformational transitions during DNA synthesis.³⁴ Nonetheless, several missense mutations in the polymerase core have been identified in XP-V patients whose structural and functional impacts remain largely uncharacterized.^{33,133,192} Computational studies have suggested that these variants may destabilize the catalytic core or interfere with DNA binding, but these hypotheses await experimental confirmation.

In this study, we investigated the structural and functional effects of six XP-V-associated Pol η variants along the catalytic cycle, before and after DNA translocation. By integrating crystallographic data and computational modeling, we aim to elucidate how these mutations perturb Pol η function and contribute to XP-V pathogenesis. This understanding will not only clarify the enzyme role in TLS but may also guide the development of therapeutic strategies targeting XP-V and related disorders. Here, we performed classical molecular dynamics (MD) simulations of the mutated Pol η in complex with a healthy DNA substrate, comparing it with the previously investigated wild-type enzyme.²⁷ We focused on the following pathogenic variants: Ala117Pro, Thr122Pro, Gly263Val, Arg93Pro, Arg111His, and Arg361Ser. These

variants were selected based on their known association with XP-V and their potential impact on Pol η function.³³ The simulations were performed employing the same protocol used for the wild-type system for consistency (section A.1).²⁷ Classical MD simulations performed for every pathogenic variant comprise three independent replicas of $\sim 1.0 \mu\text{s}$ each, for a total of $\sim 3.0 \mu\text{s}$ per variant per state. Indeed, every variant was simulated in both the pre-translocation (PRE) and post-translocation (POST) states, for a total of twelve different systems. The PRE state corresponds to the enzyme before the translocation occurs, while the POST state corresponds to the enzyme after the translocation has taken place.

5.2 Ala117Pro and Thr122Pro mutants disrupt the catalytic site

Ala117 and Thr122 are located in the palm subdomain, in proximity to the catalytic triad (Asp13, Asp115, and Glu116) and the DNA substrate (Figure 5.1). Ala117 directly interacts with Glu116, one of the catalytic residues. Substitution of Ala117 or Thr122 with proline introduces a rigid cyclic structure that can distort local secondary structure and alter the orientation of nearby residues, potentially affecting catalysis.

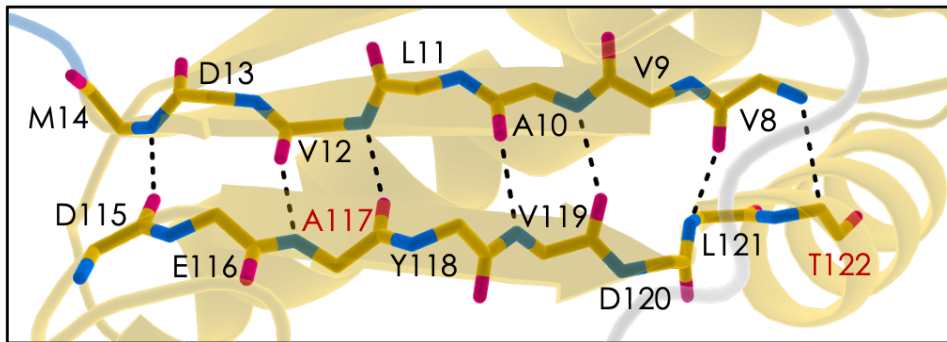


Figure 5.1: **Structural overview of Ala117 and Thr122 in wild-type Pol η :** X-ray structure of human Pol η (PDB ID: 4ECX) in complex with a DNA substrate (teal). The polymerase is shown in cartoon representation, with the finger, palm, thumb, and little-finger subdomains colored in blue, gold, pink, and red, respectively. The mutated residues studied in this work are highlighted as spheres. The 2D structure of the CPD lesion is shown in the inset (ChemDraw 25.0.2).

In the Ala117Pro mutant, we observed a marked distortion of the active site geometry, with altered orientations of the catalytic residues compared to the wild-type enzyme. This distortion likely impairs coordination of the two metal ions required for catalysis. The proline substitution induces a kink in the local backbone, disrupting the positioning of key residues in the catalytic triad. Hydrogen-bond analysis of the two antiparallel β -strands (residues 8-14 and 115-122) forming the catalytic core revealed a clear reduction in interstrand hydrogen bonds (Figure C.1). In the wild type, these strands form a stable network that maintains the geometry of the active site.³³ In the Ala117Pro mutant, the loss of key interactions, particularly the Met14-Asp115 hydrogen bond, destabilizes the catalytic triad. These residues normally coordinate the Mg^B ion, and their separation prevents proper cation binding. Even in simulations without magnesium, the wild-type triad coordinated Na^+ ions from the solvent, whereas this coordination was lost in the mutant.²⁷ Additional weakening of the Cys16-Asp20 and Val12-Pro117 interactions further destabilizes the catalytic site and likely impairs enzymatic activity.

Thr122 lies at the N-terminus of an α -helix directly connected to the catalytic center. The Thr122Pro substitution induces similar structural perturbations, propagating through the catalytic region.³³ As with Ala117Pro, hydrogen bonding between the same β -strands is reduced (Figure C.2), with loss of stabilizing interactions such as Ala10-Val119 and Val8-Leu121. This weakening increases local flexibility and destabilizes the catalytic core, ultimately compromising Pol η activity.

Together, these results suggest a shared mechanism for Ala117Pro and Thr122Pro: both mutations disrupt the hydrogen-bond network that stabilizes the catalytic triad, leading to misalignment of active-site residues and impaired coordination of catalytic metal ions.

5.3 Gly263Val, Arg93Pro, and Arg111His affect translocation and DNA binding

Gly263, Arg93, and Arg111 are located in the thumb, finger, and palm subdomains of Pol η , respectively, and play key roles in maintaining the correct positioning of the DNA substrate during synthesis and translocation (Figure 5.2). Alterations in these residues can perturb enzyme-DNA interactions, affecting both processivity and fidelity.

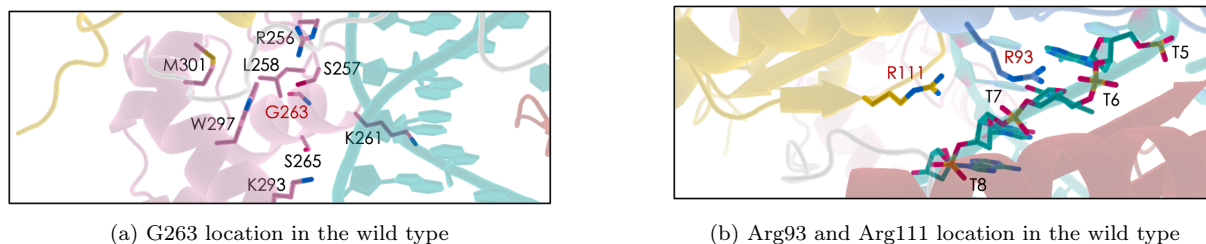


Figure 5.2: **Structural overview of Gly263, Arg93, and Arg111 in wild-type Pol η :** (Left) Gly263 is located in the thumb subdomain (pink) and interacts with the primer strand of the DNA substrate (teal). (Right) Arg93 and Arg111 are located in the finger (blue) and palm (gold) subdomains, respectively. Despite being in different subdomains, Arg93 and Arg111 are spatially close and both interact with the template strand of the DNA substrate, acting as a molecular splint.³³

Gly263Val. Gly263 lies within the thumb subdomain, which stabilizes the DNA duplex during translocation. The substitution of glycine with valine introduces a bulkier side chain, creating steric hindrance that perturbs the local conformation of the thumb and reduces enzyme-DNA contacts. Gly263 is positioned near Trp297, a residue involved in orienting the incoming nucleotide.¹⁹³ Consequently, the Gly263Val mutation may misalign this region, affecting nucleotide binding and translocation.³³ In simulations of the Gly263Val variant, we observed a reduced number of enzyme-DNA contacts and frequent detachment of the thumb subdomain from the nucleic acid. This unbinding hinders the proper anchoring of the DNA, compromising translocation efficiency. Interestingly, a similar opening of the thumb subdomain was also observed in the wild type when an inadequate collective variable (CV) was used to describe translocation, indicating that thumb detachment represents an unproductive conformational state.²⁷ Therefore, the Gly263Val mutation likely promotes this unfavorable open state, impairing DNA translocation and processivity.

Arg93Pro. Arg93, located in the finger subdomain, interacts with the phosphate backbone of the template strand, stabilizing its position during nucleotide incorporation. The Arg93Pro mutation replaces a flexible, positively charged residue with a rigid proline, which introduces a kink in the local helix (Figure C.3) and disrupts DNA contacts. This alteration weakens the molecular splint that preserves the B-form DNA conformation.³³ Despite this disruption, the Arg93Pro mutation does not significantly affect the catalytic core: the triad (Asp13, Asp115, Glu116) remains properly aligned, and the hydrogen-bond network between the β -strands (residues 8-14 and 115-122 in PDB ID: 4ECX) is largely preserved. Thus, Arg93Pro specifically impairs DNA binding and translocation without altering catalysis. Moreover, the kink induced by proline affects the orientation of Tyr92, a residue essential for the steric-gate mechanism that prevents ribonucleotide incorporation (Figure C.4).^{44, 46–48, 51, 194, 195} This suggests that Arg93Pro may compromise the enzyme fidelity and TLS efficiency.

Arg111His. Arg111, positioned in the palm subdomain, contributes to DNA stabilization through interactions with the template strand and cooperates with Arg93 to form a molecular splint. Substitution with histidine alters both charge and hydrogen-bonding capability, potentially modulating DNA interactions. We performed molecular dynamics simulations considering multiple protonation states of His111 (δ , ϵ , and fully protonated). Although pKa calculations (~ 4.5) suggest a predominantly deprotonated form, both the protonated and δ tautomers showed higher stability than the ϵ form. In these two stable states, His111 formed a π -stacking interaction with Pro316 (in the linker region), maintaining a geometry that allows dynamic flipping between contacts with Ser96 or the DNA template, similar to the wild-type Arg111.²⁷ This flexibility appears crucial for efficient translocation. In contrast, the ϵ tautomer failed to interact with Pro316 and instead formed a transient salt bridge with Arg313, a highly mobile residue in the linker region (Figure 5.3). This alternative interaction prevented His111 from contacting the DNA phosphate backbone (T6-T7), destabilizing the enzyme-DNA complex. Although the Arg111His mutation does not fully disrupt the enzyme structure, it modifies the equilibrium of key interactions, likely reducing translocation efficiency and overall processivity.

5.4 Arg361Ser impairs the stability of the little-finger subdomain

Arg361 is located in the little-finger subdomain, near the C-terminal region of the polymerase core. In the wild-type system (chapter 3), this residue forms a stabilizing salt bridge with Asp355, which contributes to maintaining the structural integrity of the subdomain. Although Arg361 does not directly contact the DNA substrate, it lies close to residues involved in DNA binding, such as Arg313, and helps position them correctly during translocation.³³

The Arg361Ser mutation removes the positive charge at this position, breaking the salt bridge with Asp355 and destabilizing the little-finger subdomain (Figure 5.3). Molecular dynamics simulations revealed that this substitution increases the flexibility of the subdomain and perturbs its interactions with the DNA backbone, particularly in the post-translocation state. As a result, the enzyme exhibits reduced ability to anchor and stabilize the DNA during translocation, leading to decreased processivity. In addition, the Arg361Ser variant induces a local rearrangement: the side chain of Gln395 flips to maintain contact with Ser361, altering the conformation of the major helix at the interface between the little-

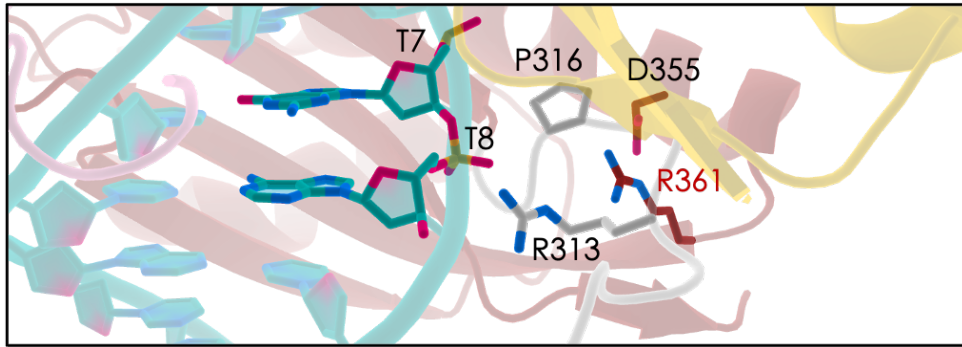


Figure 5.3: **Structural overview of Arg361 in wild-type Pol η .** Arg361 is located in the little-finger subdomain (red) and forms a stabilizing salt bridge with Asp355. The polymerase is shown in cartoon representation, with the palm and little-finger subdomains colored in gold and red, respectively.

finger and finger subdomains. This region, rich in positively charged residues crucial for DNA binding and translocation, becomes less compact and more dynamic. Altogether, these observations indicate that the Arg361Ser mutation disrupts both inter- and intra-domain electrostatic networks, impairing the coordination of the translocation process.

5.5 Discussion

The analysis of the XP-V variants Gly263Val, Arg93Pro, Arg111His, and Arg361Ser highlights distinct yet converging effects on the structural integrity and translocation mechanism of human DNA polymerase η (Pol η). Despite being located in different subdomains, all these mutations ultimately impair the enzyme ability to stably interact with the DNA substrate, thus reducing processivity and potentially compromising lesion bypass efficiency.

The Gly263Val substitution, positioned in the thumb subdomain, alters local packing and introduces steric hindrance that prevents proper anchoring of the DNA duplex. This perturbation destabilizes the thumb-DNA interface and promotes premature subdomain opening, a conformational change also observed under suboptimal translocation conditions in the wild-type system.²⁷ Notably, Gly263 is located near an allosteric binding pocket involving Trp297, a residue critical for orienting the incoming nucleotide.¹⁹³ By misaligning this region, the Gly263Val mutation may interfere with nucleotide binding as well as DNA translocation.³³ These findings suggest that the Gly263Val variant mimics a dysfunctional translocation intermediate, in which loss of thumb-DNA contacts and altered allosteric interactions hinder forward movement of the template-primer duplex.

Arg93Pro and Arg111His, though located in the finger and palm subdomains, respectively, act in a coordinated manner as part of the so-called molecular splint.³³ Their mutations disrupt complementary interactions that maintain the B-form conformation of DNA during synthesis.³³ The Arg93Pro substitution abolishes key electrostatic contacts with the phosphate backbone and induces a local helical kink that displaces Tyr92, a residue involved in nucleotide discrimination through the steric-gate mechanism.^{44, 46–48, 51, 194, 195} Arg111His, conversely, affects translocation more subtly through altered protonation equilibria and dynamic rearrangements that weaken DNA binding. Together, these mutations compromise the fine-tuned coordination between the finger and palm subdomains required for

accurate and efficient DNA synthesis.

Finally, the Arg361Ser mutation affects the little-finger subdomain, destabilizing its architecture by breaking the salt bridge with Asp355. This leads to increased flexibility and loss of anchoring interactions with the DNA backbone. Moreover, conformational rearrangements propagated to neighboring residues, such as Gln395 and Arg313, indicate a redistribution of intra-domain contacts that may further disrupt the coordination between the little-finger and finger subdomains. Such structural perturbations are consistent with a loss of processivity and reduced ability to maintain the polymerase-DNA complex during translocation.

Overall, these mutations illustrate how distinct local perturbations can converge on a shared molecular outcome: impaired translocation and DNA binding. The structural destabilization observed across different subdomains supports the notion that Pol η relies on a delicate balance of inter-domain communication to couple nucleotide incorporation with DNA movement. Disruption of this balance, as seen in XP-V-associated variants, likely underlies the reduced fidelity and efficiency of lesion bypass, contributing to the pathological phenotype observed in patients.

Chapter VI – Conclusions and Perspectives

The aim of this work was to study the dynamical movement of a DNA polymerase (Pol) along the DNA strand. This shift, known as translocation, is a crucial step in Pol-mediated nucleic acid polymerization, a process essential for storing and transmitting genetic information in all living organisms.⁷ During this process, Pols follow a common step-based catalytic cycle comprising chemical and physical steps. Starting with DNA substrate binding, the Pol•DNA binary complex is formed. Subsequent steps involve accommodating the incoming nucleotide (dNTP) at the 3'-terminus of the growing strand (primer), often after significant conformational changes in the enzyme architecture. The nucleotide is then attached to the growing strand via nucleophile addition, including 3'-OH deprotonations for subsequent phosphoryl transfer. Once this chemical step has occurred and pyrophosphate is released, the newly elongated double-stranded DNA must shift within the Pol enzyme to recreate the initial configuration at the metal-aided reactive center, where nucleotide addition occurs.²⁷

Despite its importance, the mechanism by which the Pol-DNA complex performs translocation remains incompletely understood.⁷ Recently, time-resolved X-ray structures of human DNA polymerase η (Pol η) captured the course of DNA synthesis at atomic resolution, including structures of the Pol•DNA complex before and after translocation.³⁴ Pol η is a low-processive polymerase belonging to the Y family, capable of synthesizing DNA even past specific lesions through a two-metal-ion mechanism.⁷⁶

Using molecular dynamics (MD) simulations of the Pol η wild-type system, we characterized the metastable states involved in translocation.²⁷ While MD simulations of single-subunit T7 RNA Pol and multi-subunit RNA Pols highlight the role of a flexible α -helix in translocation, this model is not suitable for Pol η due to its unique β -strand and extended loop configuration.^{36,37} Since classical MD lacks the temporal resolution to capture the full transition pathway, we employed enhanced sampling techniques to explore regions of conformational space that are otherwise inaccessible at room temperature.^{172,196} A suitable set of collective variables (CVs) was essential to guide the system along the relevant translocation pathways. We found that DNA translocation in Pol η occurs asynchronously, following one of two pathways, each characterized by previously unknown intermediates in which one strand is post-translocated, and the other remains pre-translocation. Incorporating this information into the CVs allowed us to distinguish and sample the two pathways.²⁷

Free energy surfaces (FES) of the complete translocation process revealed a minimum energy path where the growing strand translocates before the template strand.²⁷ Analysis of transitions highlighted the importance of positively charged residues at the Pol•DNA interface, which act like wipers, grasping and sliding the DNA substrate within the enzyme cavity. The greater number of interactions between the template strand and protein explains the observed energy differences. These results have been published in the Journal of the American Chemical Society in 2025.²⁷

Although numerous studies on polymerases exist in the literature, based on standardized and widely recognized pipelines, quantifying the specific contribution of translocation remains challenging.⁵³ No direct method currently exists to separate the energetic contribution of translocation from that of binding. Consequently, experimental data on Pol η translocation are limited, and none address mutant systems. Identifying mutations capable of modulating translocation could open opportunities for engineering poly-

merases for synthetic biology and biotechnology.⁷ To explore this, we applied our computational pipeline to single-point mutations of the Pol η •DNA system, focusing on positively charged residues at the primer interface. In parallel, collaborators at the Italian Institute of Technology (IIT) developed an experimental pipeline to measure the processivity rate of the wild-type system and four mutants. Since processivity reflects contributions from both translocation and binding, computational studies are needed to disentangle the atomistic details. Comparison with the wild-type system provided new insights into the molecular mechanism of Pol η translocation.

During my PhD, I also spent a research stay at Forschungszentrum Jülich GmbH under the supervision of Prof. Paolo Carloni. This experience was highly valuable scientifically and personally, as it allowed me to work with some of the most powerful supercomputers in Europe, enabling large-scale simulations that would have been otherwise impossible. My research there focused on the pathological aspects of Pol η , particularly mutations identified in Xeroderma Pigmentosum Variant (XP-V) patients.^{13,33,191} This direction aligned with the mission of the Institute of Computational Biomedicine, which develops and applies molecular simulations, machine learning, and bioinformatics tools to study complex biomedical problems, including neurotransmission in the human brain. The stay in Jülich broadened my scientific skills, exposed me to diverse methodological perspectives, and fostered interdisciplinary collaborations, representing a key step in the progress of my PhD and my development as an independent researcher.

Looking forward, I plan to further strengthen this collaboration by applying the institute's expertise in hybrid QM/MM methodologies to study Pol η , investigating how pathological mutations affect catalytic activity at the electronic level, complementing the large-scale MD studies conducted during my PhD.¹⁹⁷ To support these projects, I have recently submitted a computational proposal to access the institute high-performance supercomputers, ensuring the resources required. These investigations have the potential to deepen our understanding of Pol η -related diseases, such as XP-V, and open new perspectives for biomedical applications.

Appendix A: Supplementary Material for Chapter 3

*Coordinated Residue Motions at the Enzyme–Substrate Interface
Promote DNA Translocation in Polymerases*, published in
Journal of American Chemical Society (2025).

Alessia Visigalli¹, Enrico Trizio², Luigi Bonati², Pietro Vidossich¹, Michele Parrinello^{*,2},
and Marco De Vivo^{*,1}.

¹ Laboratory of Molecular Modeling and Drug Discovery, Istituto Italiano di
Tecnologia, Via Enrico Melen 83, 16142, Genoa, Italy

² Atomistic Simulations, Istituto Italiano di Tecnologia, Via Enrico Melen 83, 16142,
Genoa, Italy

* Corresponding authors:

Marco De Vivo, marco.devivo@iit.it

Michele Parrinello, michele.parrinello@iit.it

Supporting Information

A.1 Material and Methods

A.1.1 Structural Models Preparation.

Pre-translocation state of Pol η simulated in this work was modelled from the X-ray structure obtained after 300s of magnesium incubation (PDB ID: 4ECX, 1.74Å).³⁴ Due to the transient nature of the transition state, this structure comprises both reactants (Pol η ·DNA·dATP) and products (Pol η ·DNA·PPi) states of catalysis with different occupancies (30% for the reactants and 70% for the products). Post-translocation state structure of Pol η was modelled from the structure corresponding to normal-translocation state (PDB ID: 4ED8, 1.52Å).³⁴ Here, we replaced the TG mismatch present in the DNA sequence with the correct TA, in order to match the DNA sequence of the two models. Both models comprise a binary complex with protein and DNA; thus, we removed the ligands and the ions present in the active site (Ca²⁺ and Mg²⁺). The two systems were then solvated in an orthorhombic box with a 12 Å layer of TIP3P water molecules.¹⁴⁷ Counterions Na⁺ and Cl⁻ were added to neutralize the system and reach the physiological ionic strength of 0.15 M using tleap package of AmberTools22.¹⁹⁸ Overall, each system comprises $\sim 70,000$ atoms, including the protein, DNA, water molecules and ions. The topology and the coordinates of the systems were generated using tleap.¹⁹⁸ We considered the ff14SB force field for the protein,¹⁴⁰ the bsc1 force field for the DNA.^{142,143} Monovalent ions were described using Li-Merz parameters, and the TIP3P model for waters.^{147,150}

A.1.2 Classical MD Simulations.

We performed classical MD simulations using the GPU version of the pmemd module of Amber20.¹⁹⁹ Long-range electrostatic interactions were treated using the particle mesh Ewald (PME) method with a cutoff of 10 Å.²⁰⁰ The SHAKE algorithm was applied to constrain all bonds involving hydrogen atoms, allowing us to use a 2 fs time step.²⁰¹ Periodic boundary conditions in the three dimension of the Cartesian space were applied to the systems.

System minimization and equilibration. Our simulation protocol comprises four steps of minimization and equilibration, to ensure the correct energy relaxation of the system. (I) The first minimization step was performed to minimize the reconstructed areas of the system. Indeed, the atoms reconstructed with MODELLER or tleap are free to move, while we fixed those present in the X-ray structure (see Structural and Model Preparation section).²⁰² In particular, we applied a harmonic restraint of 100 kcal mol⁻¹ Å⁻² for 10,000 steps, using the steepest descent algorithm, followed by a conjugate gradient algorithm for 10,000 steps. (II) The second step is the solvent dynamic. This step is needed to relax the solvent and the ions, allowing them to penetrate in the system cavities. We performed NPT solvent dynamics for 20 ns with 2 fs time step, using a Langevin thermostat with a collision frequency of 2 ps⁻¹ and a Monte Carlo barostat with a target pressure of 1 atm.^{153,159} (III) The third step is the side chain minimization. In this step, we applied a harmonic restraint of 100 kcal mol⁻¹ Å⁻² to the protein and DNA backbone, allowing the side chains and the solvent to equilibrate maintaining the backbone stable, with the same minimization parameters as in the first step. (IV) The fourth step is the full system minimization. In this step, we removed all the harmonic restraints applied to the system, allowing the

complete structure relaxation. All the other parameters were the same as in the previous step.

Thermalization steps. After the minimization steps, we performed a thermalization step to bring the system to the target temperature of 310 K. This is done in three stages. (I) First, we maintain the system at a fixed temperature of 100 K for 100 ps, allowing it to relax at low temperature without altering the structure. Here, we simulate in NVT ensemble, fixed the protein and DNA backbone with a harmonic restraint of $100 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$. The Langevin thermostat was used with a collision frequency of 5 ps^{-1} .¹⁵³ (II) Then, we gradually increase the temperature from 100 K to 310 K in 2 ns. This is done in NPT ensemble, with the same restraint on the protein and DNA backbone as in the previous step. The Langevin thermostat was used with a collision frequency of 5 ps^{-1} and the Monte Carlo barostat was applied to maintain the pressure at 1 atm.^{153,159} (III) Finally, we maintain the system at a fixed temperature of 310 K for 3 ns, allowing it to equilibrate at the target temperature without any restraints. Again, we simulate in NPT ensemble, where the Langevin thermostat was used with a collision frequency of 1 ps^{-1} and the Monte Carlo barostat was applied to maintain the pressure at 1 atm.^{153,159}

Production runs. After the thermalization steps, we performed the NPT production run at 310 K and 1 atm. The Langevin thermostat was used with a collision frequency of 1 ps^{-1} and the Monte Carlo barostat was applied to maintain the pressure at 1 atm.^{153,159} The production runs were performed for $2 \mu\text{s}$ for each system, using a time step of 2 fs and a cutoff of $10 \text{ \AA}$ is applied for the calculation of Lennard-Jones and Coulomb interactions alike. In total, we performed three replicas of $2 \mu\text{s}$ classical MD simulations for each system, resulting in $6 \mu\text{s}$ of simulation time for each state.

A.1.3 Enhanced Sampling Simulations.

To promote the translocation process, we performed enhanced sampling simulations using the On-the-Fly Probability Enhanced Sampling (OPES) method (subsection 2.2.3).¹⁶⁷ To do that, we first trained a machine learning CV using Deep learning algorithms (section 2.3). The fundamental step to create a good CV is the choice of the descriptors to use as input for the neural network.

Choice of descriptors. In our case, we focused on the distances between the enzyme and the DNA substrate, which are expected to change during the translocation process. First, we calculated the native contacts between the enzyme and the DNA substrate using `cpptraj`.²⁰³ The ~ 350 resulting contacts were then filtered to select only the distances between the DNA and the enzyme that are expected to change during the translocation process (Figure A.9). In particular, we selected (I) the distances present for at least the 50% of the simulation time in the pre- and post-translocation states; (II) the distances that show a significant change (more than $3 \text{ \AA}$) between the pre- and post-translocation states; and finally (III) the distances that are not correlated with each other. This resulted in a total of 69 distances between the DNA and the enzyme in both states, which were used as input for the neural network.

Deep-LDA CV. To create the Deep-LDA CV, we used the 69 selected distances as input features for a deep learning model.¹⁶⁹ The model was trained to distinguish between the pre- and post-translocation states based on these features. For the training, we used the Deep-LDA implementation available in the

mlcolvar library.¹⁷² The neural network architecture used for the training was [69, 30, 20, 15, 10, 5] nodes/layer using the rectified linear unit (ReLU) activation function. The training was performed for 500 epochs with a learning rate of 0.0005, an L_2 regularization of 10^{-4} , a within-scatter regularization S_W of 0.05, and a Lorentzian regularization of 40. This resulted in a deep-learning CV (Deep-LDA CV) where the PRE and POST states were mapped at 1 and -1, respectively.

Multitask CV. To properly describe the two translocation pathways and includes the intermediate states, we trained a 2-dimensional CV using a multitask framework (Figure 2.3.2).¹⁷² This approach allows us to include in the neural network input the data coming from both classical MD simulations in the four states (pre- and post-translocation, INT1 and INT2), both also data collected in the previous OPES simulations. In details, this framework allows the optimization of a single set of CVs, leveraging information from datasets of different sources. To do so, using as inputs the same descriptors as for the Deep-LDA CV, we expressed our CVs as the latent space of an autoencoder neural network, which was optimized to achieve two different tasks: dimensionality reduction of the transition pathways (unsupervised learning, see Figure 2.3.2) and classification of the different metastable states (supervised learning, see Figure 2.3.2).¹⁷² For the training, we used the Multi-Task CV framework available in the mlcolvar library (Figure 2.4).¹⁷² The architecture of the encoder network of the model was [69, 50, 30, 15, 5, 2] nodes/layer and symmetric for the decoder network [2, 5, 15, 30, 50, 69] nodes/layer and the shifted softplus activation was used for both networks. The training was performed for 500 epochs with a learning rate of 0.001. For the TDA part of the loss, we set the centers of the target Gaussians associated with the states to $\mu_{tg}=([-6,-6], [-3,3], [3,-3], [6,6])$ and their sigmas to $\sigma_{tg}=([0.2,0.2], [0.2,0.2], [0.2,0.2], [0.2,0.2])$, for the PRE, INT2, INT1, POST states and the two components CV1 and CV2, respectively. This resulted in the two-dimensional MLCVs, composed by CV1 and CV2 in which the states PRE, INT1, INT2, POST, were mapped at $[-6,-6],[3,-3], [-3,3], [6,6]$. The resulting CV spatially separates the two translocation pathways, allowing us to distinguish between the two asynchronous steps.

OPES Simulations. In this work, we first performed OPES EXPLORE on the one-dimensional Deep-LDA CV, collecting a large set (>80) of exploratory runs (Equation 2.2.3). At this stage, we set a barrier of $BARRIER = 20$ kcal/mol and a kernel deposition rate of $STRIDE = 50000$. In addition, to prevent the system from escaping the physical region of the CV (i.e., $\sim [-1, 1]$), we set with PLUMED two repulsive walls at $Deep-LDA\ CV \geq 1.5$ and $Deep-LDA\ CV \leq -1.5$ with force constants of $40000\text{ kcal mol}^{-1}\text{ \AA}^{-2}$.²⁰⁴ We then performed a second set of OPES simulations using the two-dimensional Multi-Task CV (CV1, CV2), observing ~ 20 in $1.2\text{ }\mu\text{s}$ over six independent replicas. Notably, our Multi-Task CV pushed the system through either of the two pathways, reaching the expected final state. For these simulations, we set the barrier to $BARRIER = 20$ kcal/mol, the deposition rate as $STRIDE = 500$, and the initial width of the kernels to $SIGMA = 0.1$. Also in this case, to prevent the system from escaping the physical region of the CV (i.e., $\sim [-6, 6], [-6, 6]$), we set with PLUMED two repulsive walls at $[-7.5, 7.5], [-7.5, 7.5]$ with force constants of $[40000, 40000]\text{ kcal mol}^{-1}\text{ \AA}^{-2}$. In addition, to prevent breaking the hydrogen bonds between the DNA bases, we defined a set of repulsive walls at $4.5\text{ \AA}$ with a force constant of $150\text{ kcal mol}^{-1}\text{ \AA}^{-2}$ for each distance involved in those hydrogen bonds.

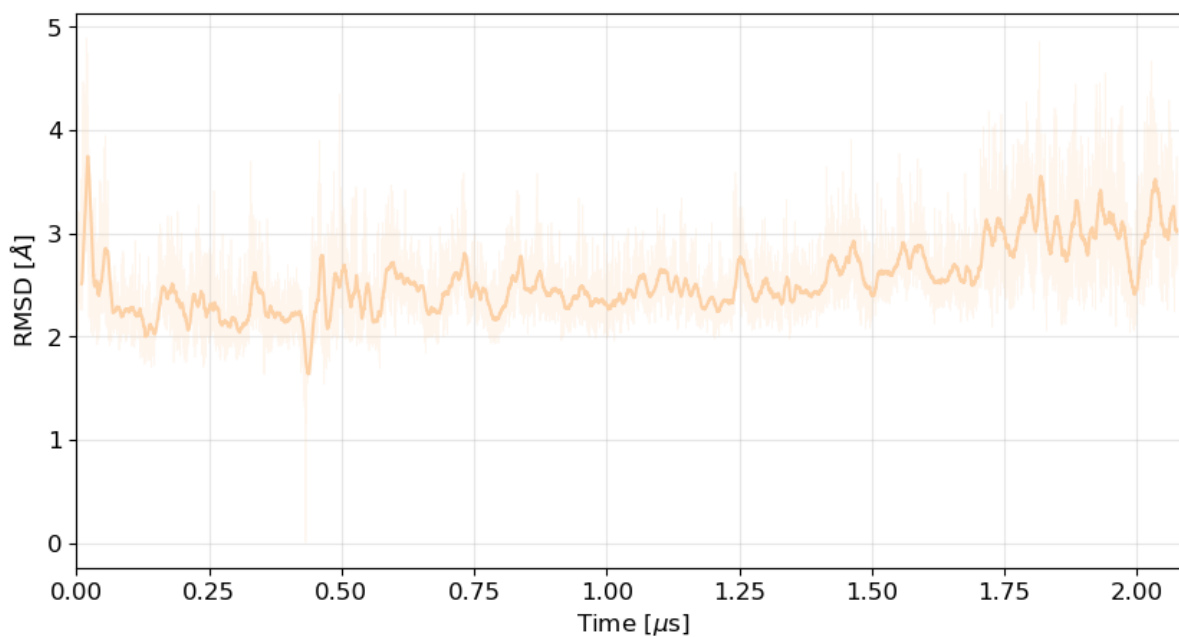


Figure A.1: **Time evolution of the Root-Mean-Square Deviation (RMSD) of the heavy atoms in the pre-translocation state over a 2 μ s simulation.** The RMSD was calculated relative to the reference structure obtained from the first cluster of a dpeak clustering analysis.^{184,205} All frames were aligned to the backbone atoms of the palm subdomain, excluding the loop regions. The solid pink line represents the rolling average of the RMSD using a window size of 100 frames, while the shaded area displays the complete dataset, highlighting the extent of structural fluctuations throughout the simulation. The gradual increase of the RMSD value around $\sim 1.70 \mu$ s is the starting point of a structural rearrangement of the system or a transition toward a different conformational state.

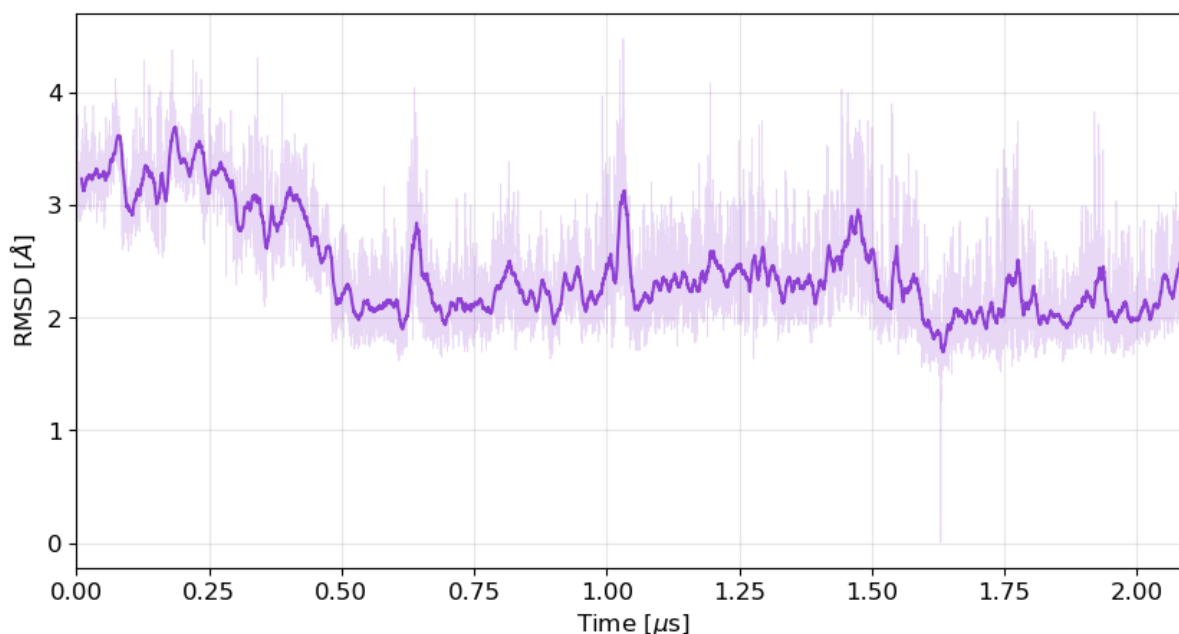


Figure A.2: **Time evolution of the Root-Mean-Square Deviation (RMSD) of the heavy atoms in the post-translocation state over a 2 μ s simulation.** The RMSD was calculated relative to the reference structure obtained from the first cluster of a dpeak clustering analysis.^{184, 205} All the frames were aligned to the backbone atoms of the palm subdomain, excluding the loop regions. The solid purple line represents the rolling average RMS using a window of 100 frames, while the shaded area displays the complete dataset, highlighting the extent of structural fluctuations throughout the simulation. The stabilization around ~ 2.2 Å suggests the system reaches a relatively stable conformation, with transient spikes indicating brief structural fluctuations or conformational changes.

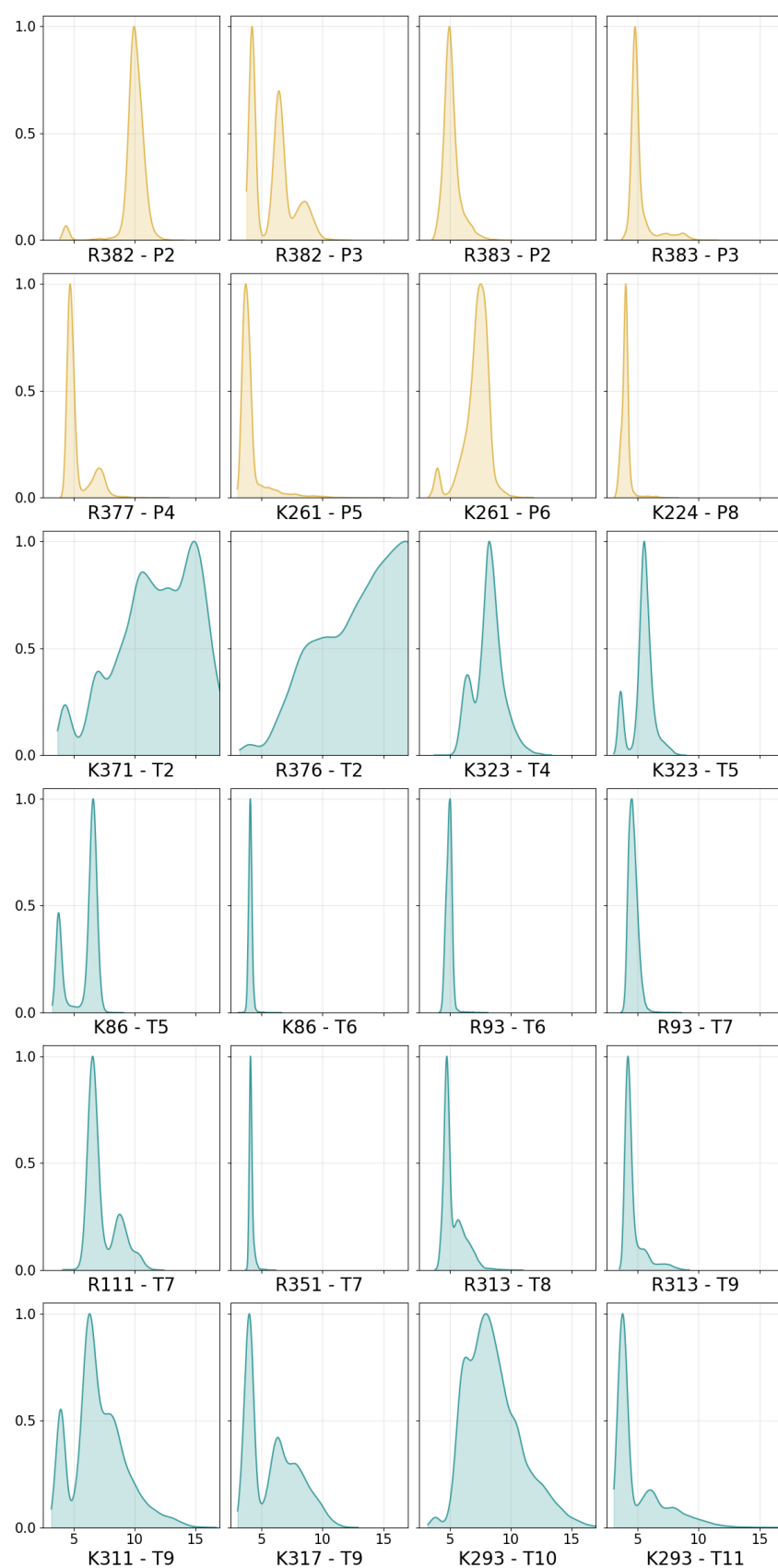


Figure A.3: **Probability distributions of key residue-DNA distances in the pre-translocation state.** Distances for primer-side interactions (gold) and template-side interactions (teal) are shown. All the distances are calculated with respect to the phosphate atom of the DNA substrate. The distances involving a lysine are calculated using the nitrogen atom of the side chain. The distances involving an arginine are calculated using the carbon atom of the guanidinium group.

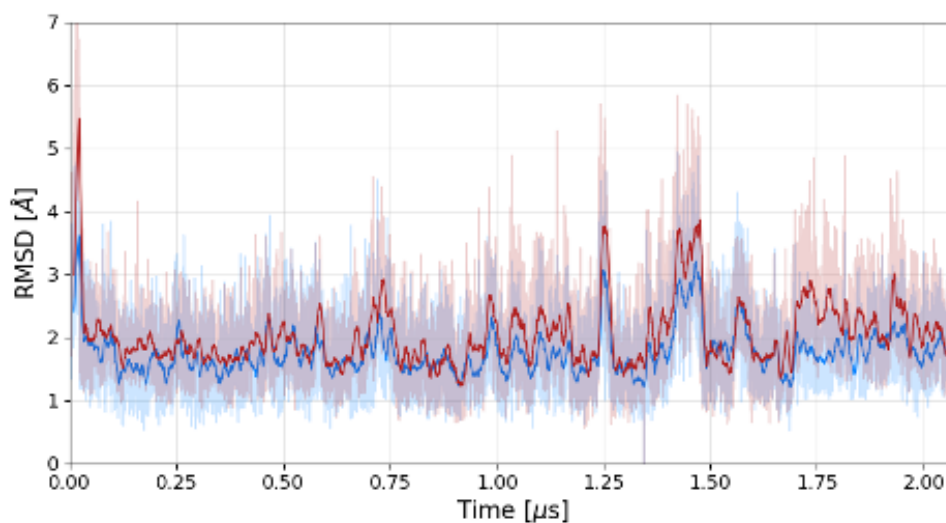


Figure A.4: **Root-Mean-Square-Deviation (RMSD) of the template strand (blue) and the primer strand (red) in the pre-translocation state.** The RMSD was calculated relative to the reference structure obtained from the first cluster of a dpeak clustering analysis.^{184,205} All frames were aligned to the backbone atoms of the palm subdomain, excluding the loop regions. The RMSD shown refers to the phosphate atoms of the double strand DNA. The solid pink line represents the rolling average of the RMSD using a window size of 50 frames, while the shaded area displays the complete dataset, highlighting the extent of structural fluctuations throughout the simulation.

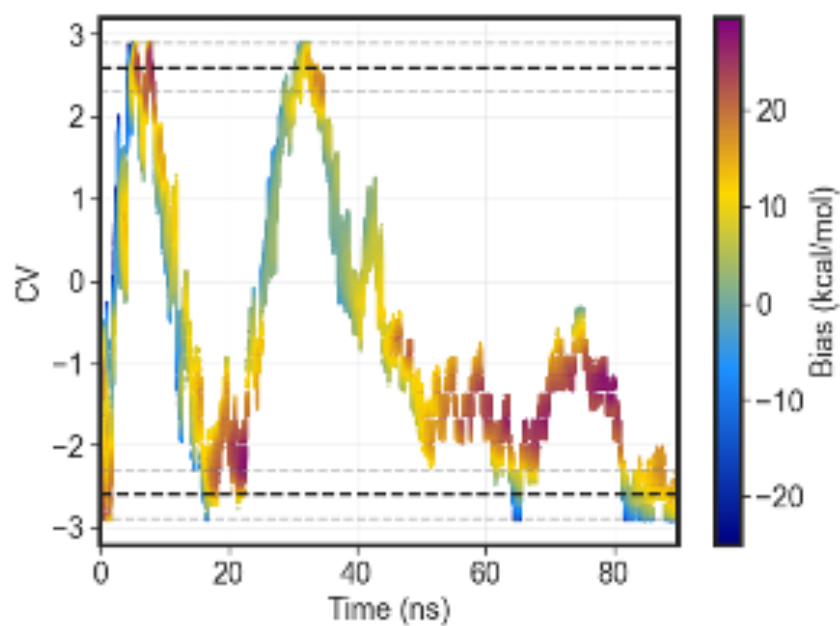


Figure A.5: **Evolution of the On-the-fly Probability Enhanced Sampling (OPES) 1D collective variable (CV) over time.**^{167,168} Here, the pre-translocation state corresponds to the CV value -2.6, while the post-translocation state corresponds to the CV value +2.6. In this example, the system was able to translocate four times in ~ 90 ns. The value set for BARRIER was 25 kcal/mol, and the color bar on the right represents the bias deposited during the simulation.

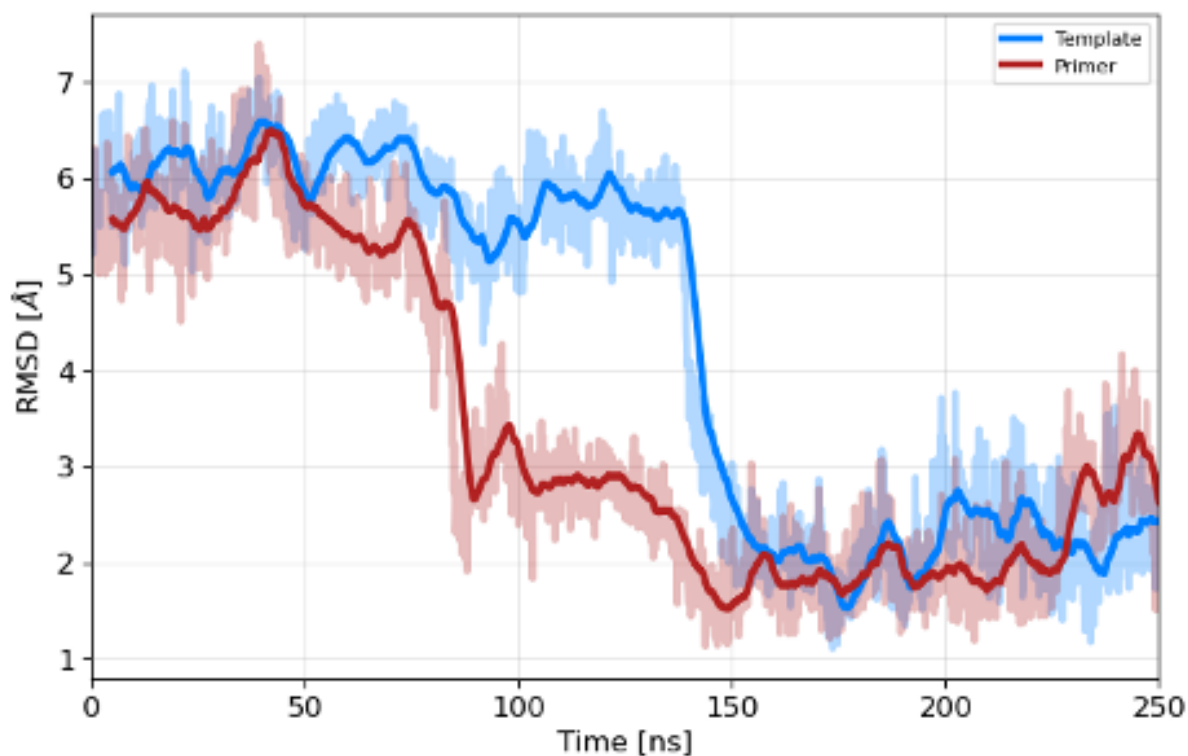


Figure A.6: **Root-Mean-Square-Deviation (RMSD) of the template strand (blue) and the primer strand (red) over time during the translocation.** In this simulation, the system was able to translocate once from the pre-translocation to the post-translocation state. The RMSD is calculated considering the dsDNA phosphate atoms with respect to the post-translocation state to appreciate the decrease in value, indicating the arrival at the correct final state. Here, the primer strand translocates ~ 50 ns before the template strand, indicating an asynchronous mechanism for DNA translocation.

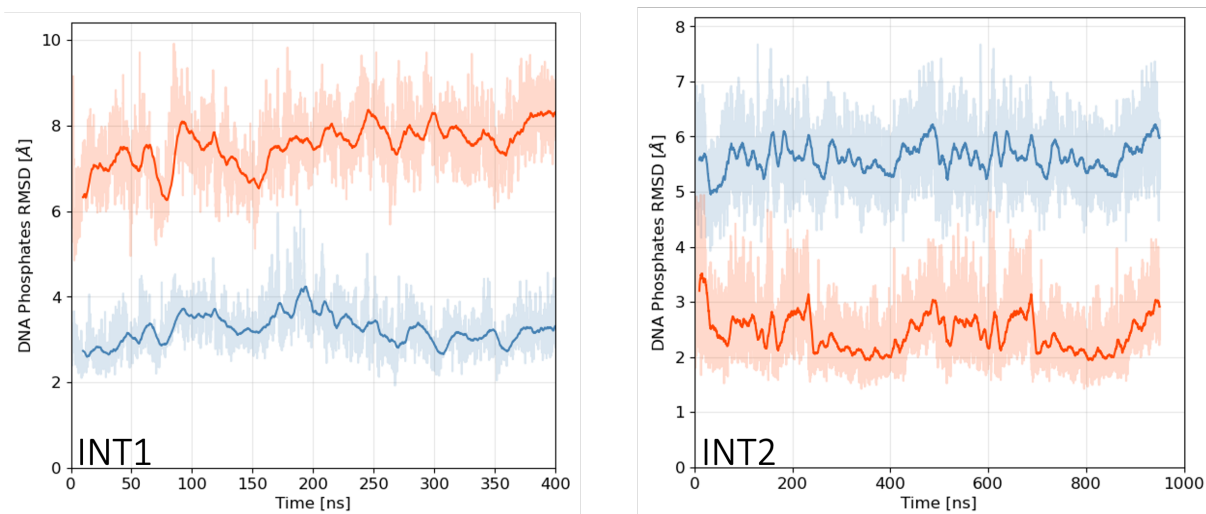


Figure A.7: **Classical MD simulations of the two intermediate states.** We calculated the RMSD of the template (blue) and the primer (red) strand of the two intermediate states INT1 and INT2, to check if they were stable. In both cases, the RMSD is calculated considering the dsDNA phosphate atoms with respect to the pre-translocation state, to appreciate the difference in value of the two strands.

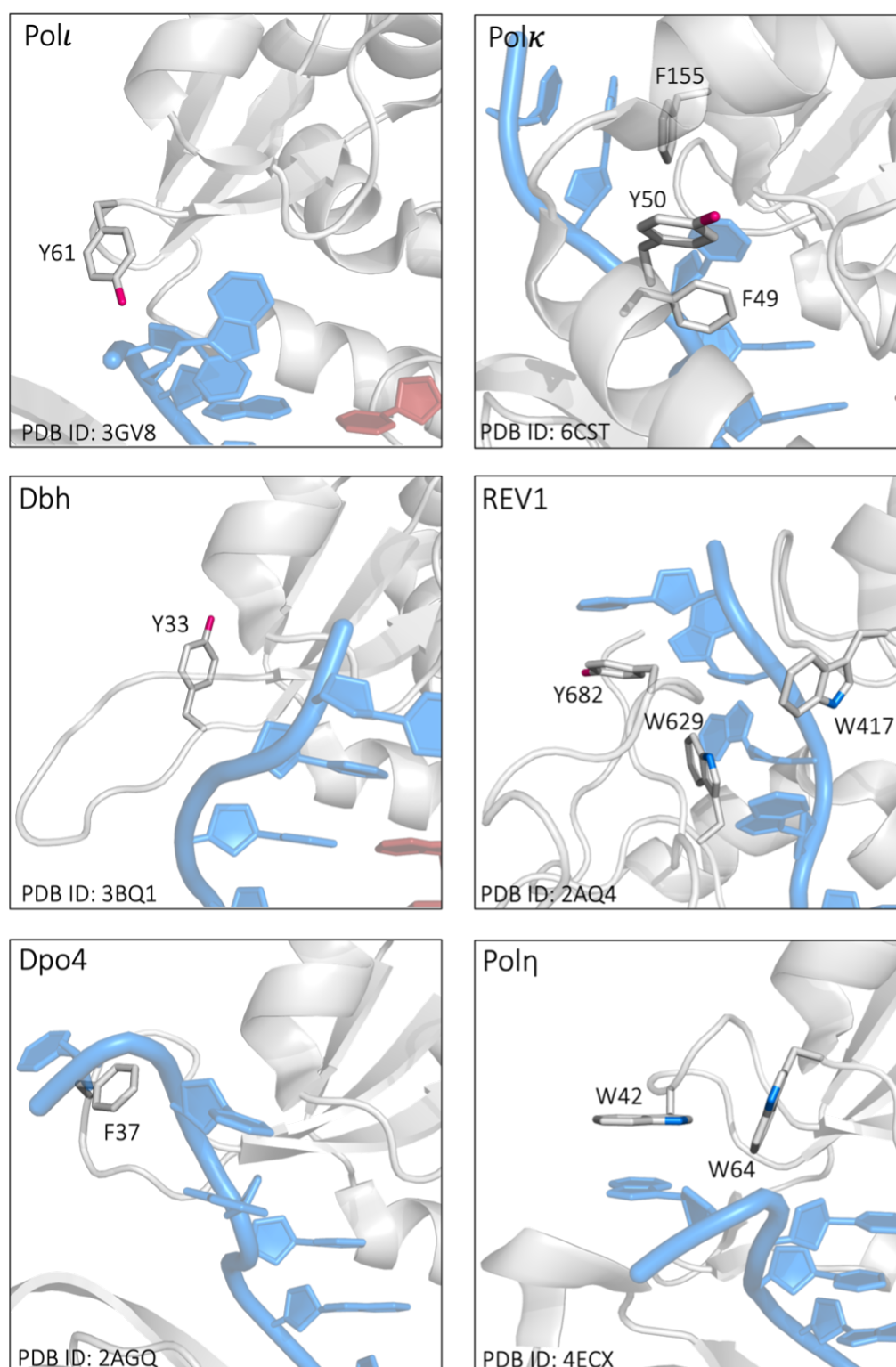


Figure A.8: **Aromatic residues at the finger•little-finger interface in Y-family Pols.** Zoom of the interface between finger and little-finger subdomain of all the six Pols belonging to the Y-family (Polι, 3GV8; Polκ, 6CST; Dbh, 3BQ1; REV1, 2AQ4; Dpo4, 2AGQ and Polη, 4ECX). In cartoon the protein (white) and the DNA substrate (the template strand in blue, the primer strand in red). In sticks, we underlined the aromatic residues present at this interface. In some cases, they can form stacking interactions with the template strand.

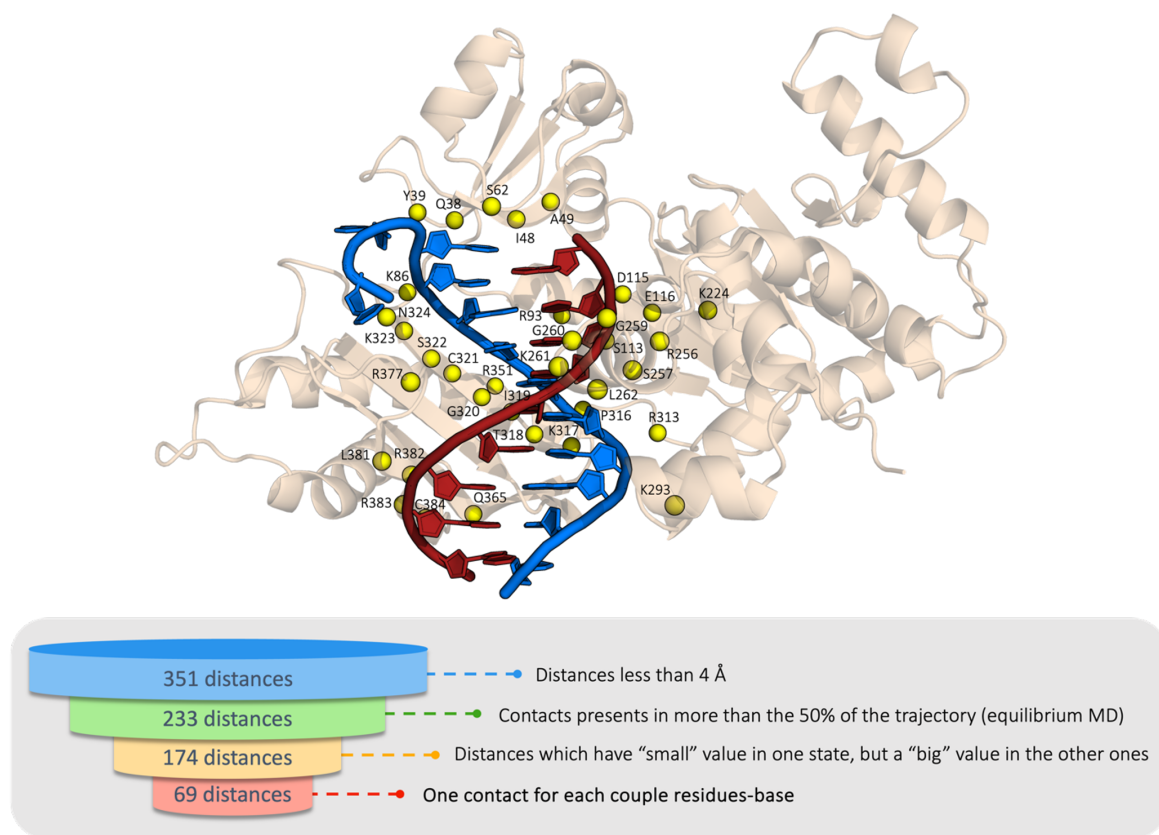


Figure A.9: Selection of descriptors for the machine learning CVs. The figure above illustrates the interaction network between the protein (light beige cartoon) and the nucleic acid (template in blue and primer in red), with yellow spheres representing the residue-base contacts selected as descriptors. The inset diagram summarizes the descriptor selection process: 351 distances correspond to all protein-DNA distances less than 4 Å. From these, 233 distances are retained as they are present in more than 50% of the equilibrium MD trajectory. A subset of 174 distances is identified, characterized by having “small” values in one state and “large” values in the other. Finally, 69 distances are selected, representing one contact for each residue-base pair.

Appendix B: Supplementary Material for Chapter 4

Hybrid Workflow to Understanding Translocation Mutant Systems of Polη.

Alessia Visigalli¹, Elisa Romeo², Natasha Margaroli², Angela Andonaia², Pietro Vidossich¹, Stefania Girotto^{*,2} and Marco De Vivo^{*,1}.

¹ Laboratory of Molecular Modeling and Drug Discovery, Istituto Italiano di Tecnologia, Via Enrico Melen 83, 16142, Genoa, Italy

² Structural Biophysics Facility, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy

* Corresponding authors:

Marco De Vivo, marco.devivo@iit.it

Stefania Girotto, stefania.girotto@iit.it

Supporting Information

B.1 Material and Methods

B.1.1 Structural Models Preparation.

Eight mutant models of human DNA polymerase η were prepared for classical MD simulations, corresponding to three alanine variants (Arg377Ala, Arg382Ala, and Arg383Ala) and the double mutant (Arg382Ala/Arg383Ala) in both pre- and post-translocation states.

All pre-translocation (PRE) models were constructed from the X-ray structure of the ternary complex (PDB ID: 4ECX, 1.74 Å),³⁴ whereas the post-translocation (POST) models were derived from the translocated binary complex (PDB ID: 4ED8, 1.52 Å). The corresponding wild-type configurations served as structural references for all comparative analyses.²⁷ The PRE model was generated using the product conformation of the 4ECX crystal (Pol η -DNA_{*n*+1} + PPi, 70% occupancy), while in the POST model the TG mismatch present in the crystal structure was replaced with a canonical TA base pair.²⁷

Active-site metal ions (Ca²⁺ and Mg²⁺) and the pyrophosphate (PPi) group were omitted, as they are not present during the translocation step.^{27,36,37} Missing loops and terminal residues were rebuilt with MODELLER 10.3, and site-directed mutations were introduced by residue substitution in the wild-type coordinates.²⁰² The stereochemical quality of the final models was verified using PROCHECK, ensuring the absence of residues in disallowed regions of the Ramachandran plot.²⁰⁶ System solvation and parameterization were performed in tleap (AmberTools24).²⁰³

B.1.2 Molecular Dynamics Simulations

All systems were equilibrated and simulated with the GPU implementation of the pmemd engine in AMBER 24.²⁰³ The ff14SB force field was used for the protein and bsc1 for the DNA duplex.^{140,142,143} Monovalent ions followed Li-Merz parameters, and solvent molecules were described with the TIP3P water model.^{147,150} Electrostatic interactions were computed with the Particle Mesh Ewald (PME) method,²⁰⁰ under fully periodic boundary conditions. A 2-fs integration time step was adopted, constraining all hydrogen-involving bonds via the SHAKE algorithm.²⁰¹ The complete minimization and equilibration protocol followed the procedure described in section A.1.

B.1.3 Clustering Analysis.

To characterize the dominant conformational states along the translocation pathway, clustering analysis was carried out in a two-dimensional “switch space” defined by the RMSD of the primer and template strands with respect to the pre- and post-translocation reference structures. For each trajectory, RMSD values were computed using phosphate atoms of the primer or template strands, sampled every 10 frames to reduce correlation between adjacent snapshots.

The four RMSD datasets were combined through a smooth switching function that continuously maps each frame onto the pre-post transition coordinate. The function is defined as:

$$S(x, y) = \frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x - x_0}{\sigma}\right) - \left[\frac{1}{2} - \frac{1}{2} \tanh\left(\frac{y - x_0}{\sigma}\right)\right], \quad (\text{B.1})$$

where x_0 represents the midpoint of the transition, corresponding to the RMSD value at which the

system shifts from pre- to post-translocation character, and σ controls the smoothness (or width) of the transition region. These parameters were chosen to balance sensitivity and numerical stability, ensuring a continuous mapping between conformational states.

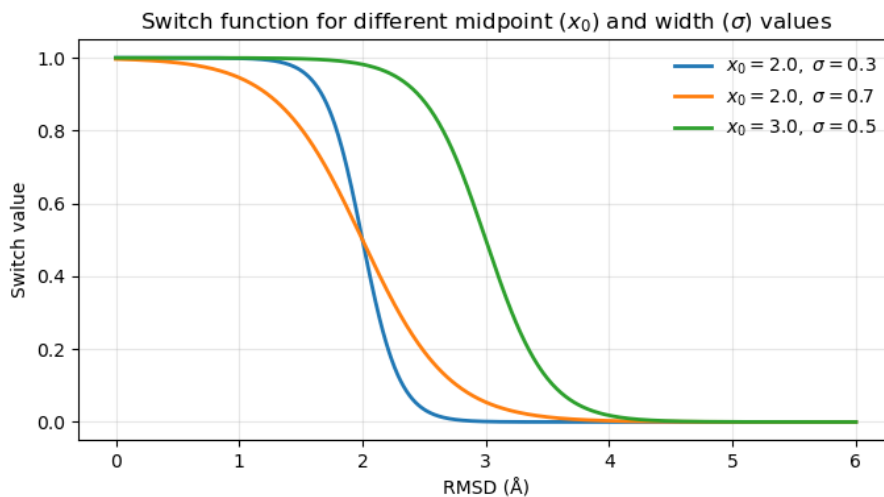


Figure B.1: **Switching Function:** Representation of the switching function used to define the two-dimensional conformational space for clustering analysis.

The resulting two-dimensional dataset (X, Y) was subjected to K-means clustering (scikit-learn implementation), testing between two and six clusters.²⁰⁷ The optimal number of clusters was selected based on the silhouette score, maximizing inter-cluster separation and intra-cluster cohesion. The centroid of the most populated cluster was extracted as the representative structure of the dominant conformational state and used as the reference for subsequent RMSD, RMSF, and structural analyses.

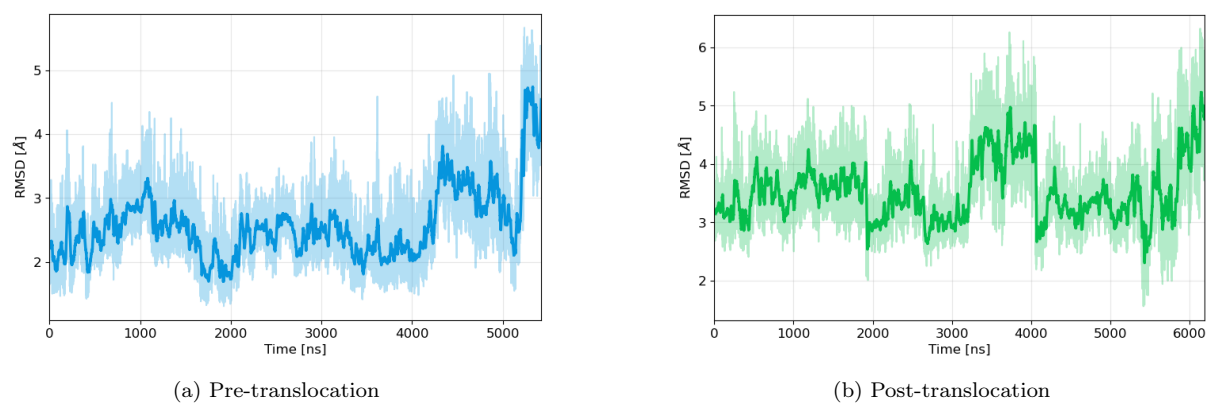


Figure B.2: **Arg377Ala RMSD.** RMSD of the protein backbone along the MD simulations for both (a) pre-translocation and (b) post-translocation systems.

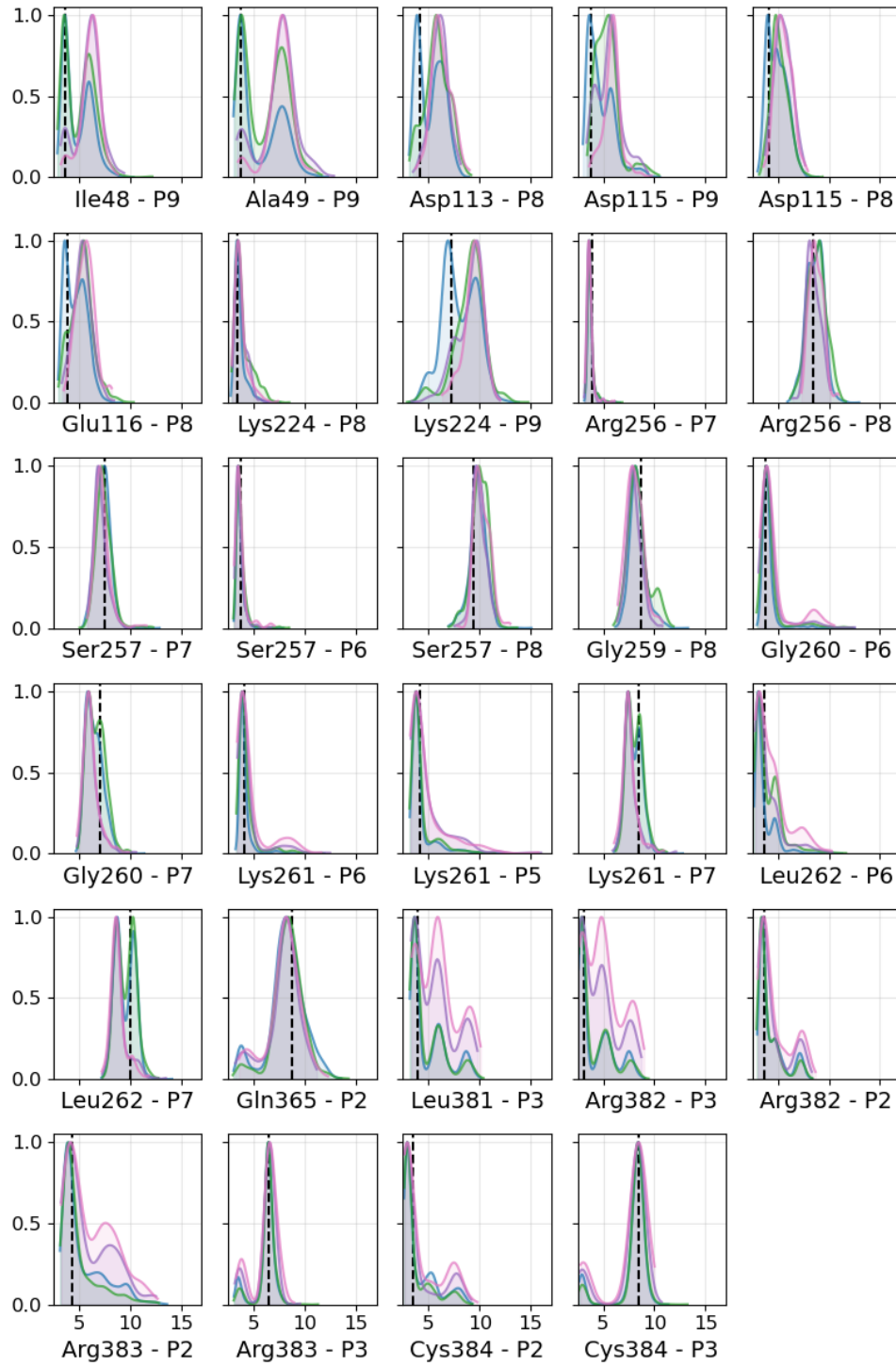


Figure B.3: **Arg377Ala descriptors primer** Density function of descriptors involving primer strand in Arg377Ala pre-translocation state.

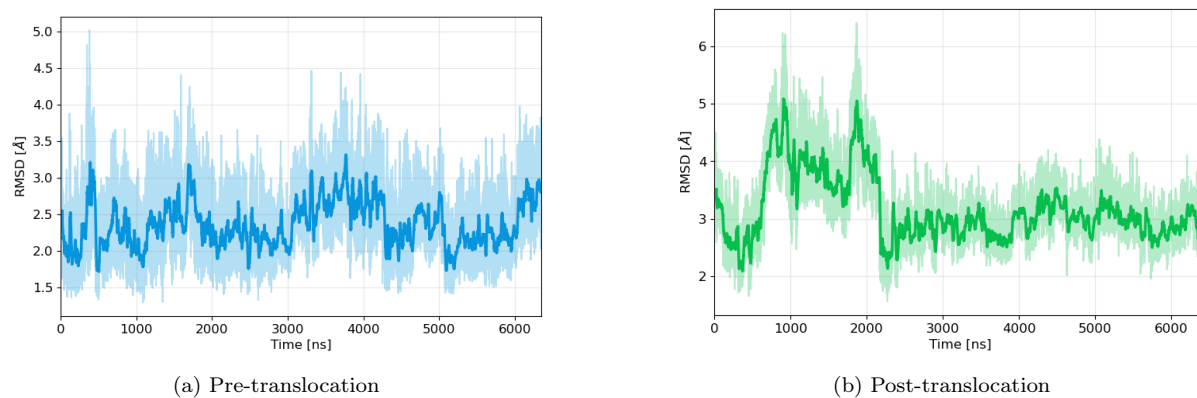


Figure B.4: **Arg382Ala RMSD.** RMSD of the protein backbone along the MD simulations for both (a) pre-translocation and (b) post-translocation systems.

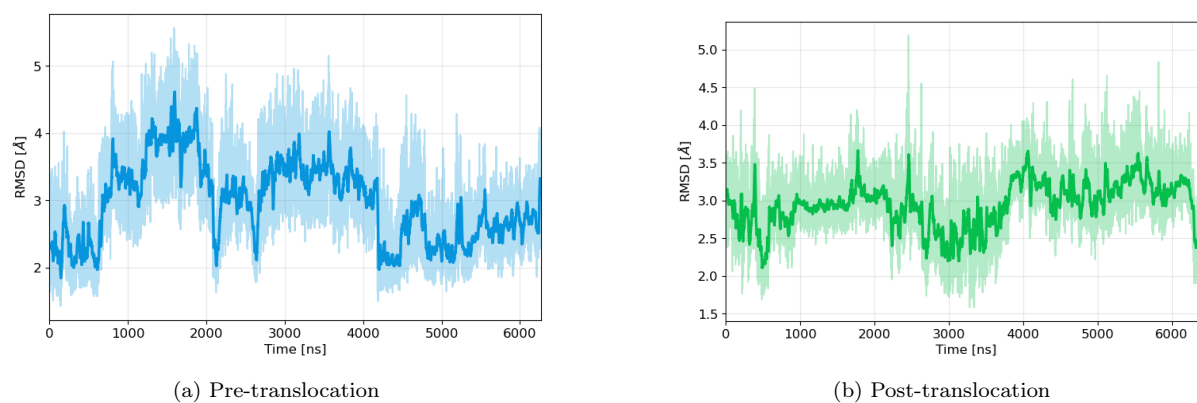


Figure B.5: **Arg383Ala RMSD.** RMSD of the protein backbone along the MD simulations for both (a) pre-translocation and (b) post-translocation systems.

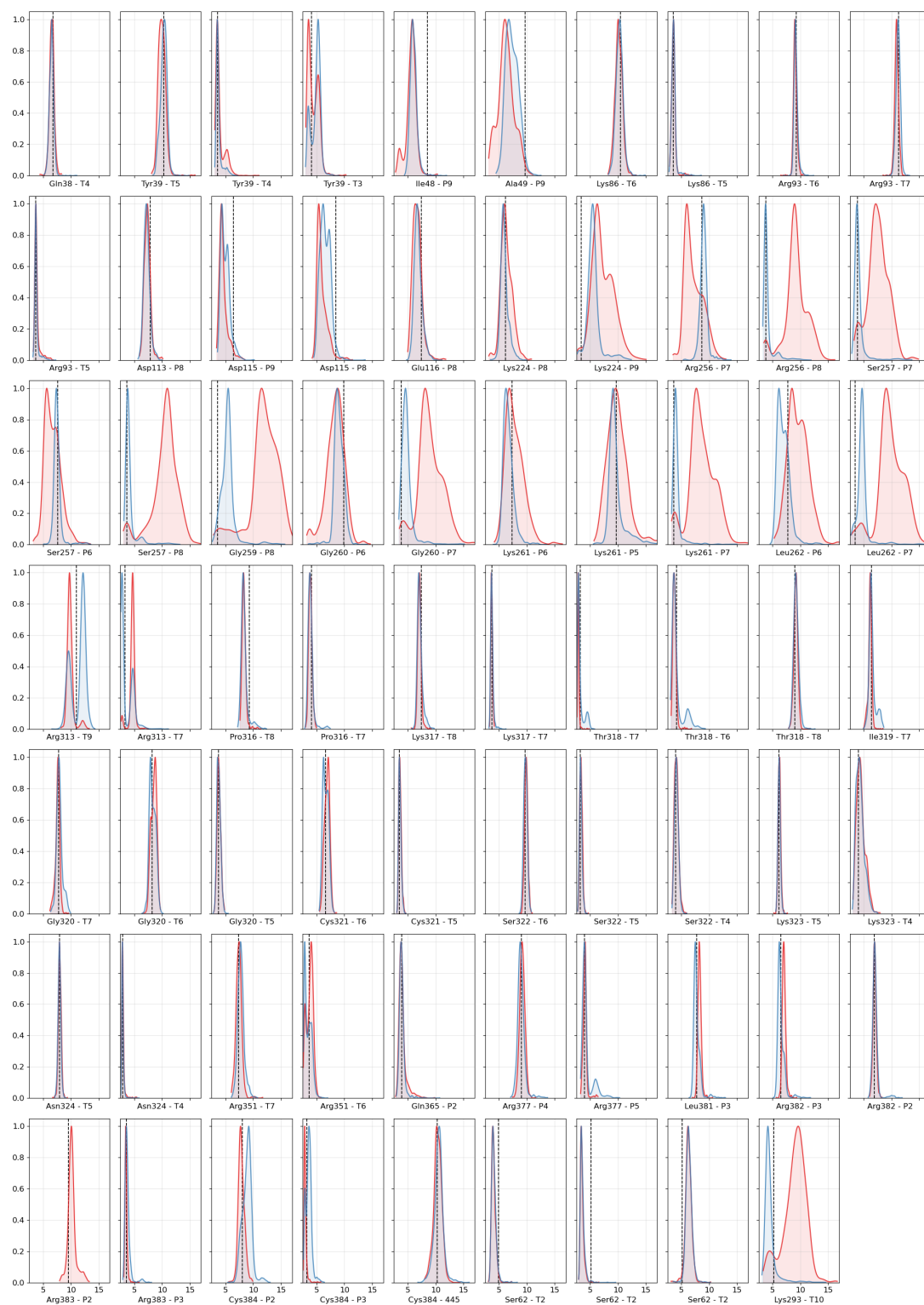


Figure B.6: **Arg382Ala/Arg383Ala** descriptors primer Density function of descriptors involving primer strand in Arg382Ala/Arg383Ala pre-translocation state.

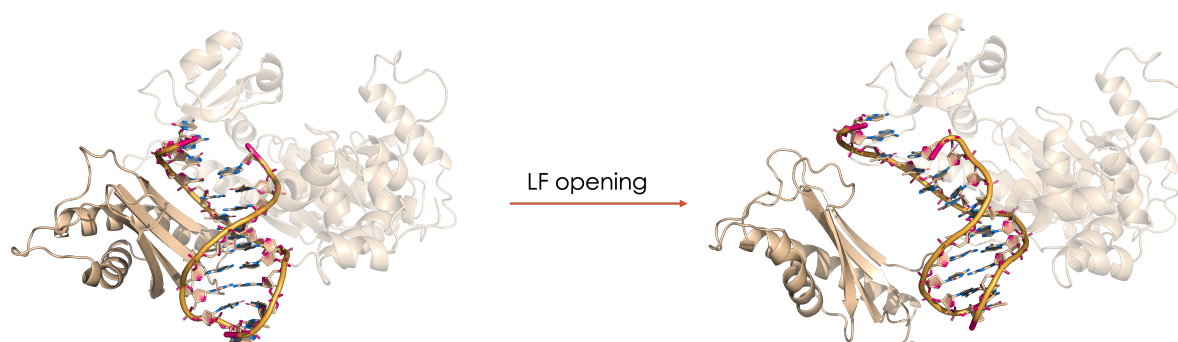


Figure B.7: **Little Finger Opening in the double mutant system** Structural representation of the little finger opening in Arg382Ala/Arg383Ala double mutant system.

Appendix C: Supplementary Material for Chapter 5

Decoding the Impact of XP-V Associated Mutations in DNA Polymerase η through Atomistic Simulations..

Alessia Visigalli¹, Paolo Carloni², and Marco De Vivo^{*,1}.

¹ Laboratory of Molecular Modeling and Drug Discovery, Istituto Italiano di
Tecnologia, Via Enrico Melen 83, 16142, Genoa, Italy

² Computational Biomedicine, Institute of Neuroscience and Medicine INM-9,
Forschungszentrum Jülich GmbH, 52428 Jülich, Germany

³ Department of Physics, RWTH Aachen University, 52056 Aachen, Germany

⁴ JARA Institute: Molecular Neuroscience and Imaging, Institute of Neuroscience and
Medicine INM-11, Forschungszentrum Jülich GmbH, 09042 Jülich, Germany

* Corresponding authors:

Marco De Vivo, marco.devivo@iit.it

Paolo Carloni, paolo.carloni@fz-juelich.de

Supporting Information

C.1 Structural Models Preparation.

Twelve mutant models of human DNA polymerase η were prepared for classical MD simulations, corresponding to six pathogenic variants (Ala117Pro, Gly263Val, Arg93Pro, Arg111His, Thr122Pro, and Arg361Ser) in both pre- and post-translocation states.

All pre-translocation (PRE) models were derived from the X-ray structure of the ternary complex (PDB ID: 4ECX, 1.74 Å),³⁴ while all post-translocation (POST) models were based on the translocated binary complex (PDB ID: 4ED8, 1.52 Å). The corresponding wild-type systems were used as reference structures for comparison.²⁷ The pre-translocation state structure contains both reactant ($\text{Pol}\eta\text{-DNA}_n + \text{dATP}$, 30% occupancy) and product ($\text{Pol}\eta\text{-DNA}_{n+1} + \text{PPi}$, 70% occupancy) conformations. The product state was used to model the PRE complex, preserving the covalent bond between the newly incorporated nucleotide and the DNA strand. For the POST complex, the TG mismatch present in post-translocation was replaced with a canonical TA base pair to match the DNA sequence used in the PRE model.²⁷

Metal ions (Ca^{2+} , Mg^{2+}) and pyrophosphate (PPi) were removed from the active site, as they are absent during the translocation step.^{36,37} Missing residues were reconstructed using MODELLER 10.3, and the pathogenic mutations were introduced within MODELLER by substituting the corresponding residues in the wild-type structures.²⁰² The resulting models were evaluated with PROCHECK to assess stereochemical quality, ensuring that no residues fell within disallowed regions of the Ramachandran plot.²⁰⁶ Final systems were completed and parameterized using `tleap` (AmberTools24).²⁰³

C.2 Simulation Protocol.

The MD simulations to equilibrate the system were performed using the GPU version of the `pmemd` module of AMBER 24. The AMBER-ff14SB force field was used for the enzyme, while the `bsc1` force field was used for the DNA substrate.^{140,142,143} Monovalent ions were described using Li-Merz parameters, and the TIP3P model for waters.^{147,150} Long-range electrostatic interactions were calculated using the particle mesh Ewald method.²⁰⁰ Periodic boundary conditions in the three directions of Cartesian space were applied. A time integration step of 2 fs was used, and the lengths of all bonds involving hydrogen atoms were constrained using the SHAKE algorithm.²⁰¹ The protocol used to equilibrate, minimize and thermalize the systems is described in detail in section A.1.

C.3 Data and Software Availability.

The used PDB files were downloaded from the RCSB Protein Data Bank (<https://www.rcsb.org/>). MODELER version 10.3 was used to model the two states, adding the missing residues to the protein (<https://salilab.org/modeller/>). `tleap` from AmberTools24 was used to create the topology of the system and to add the missing DNA bases. The GPU version of `pmemd` from AMBER24 was used to perform classical MD simulations (<https://ambermd.org/>). All the structures in the figures have been produced using the educational version of PyMOL 3.1 (<https://www.pymol.org>).

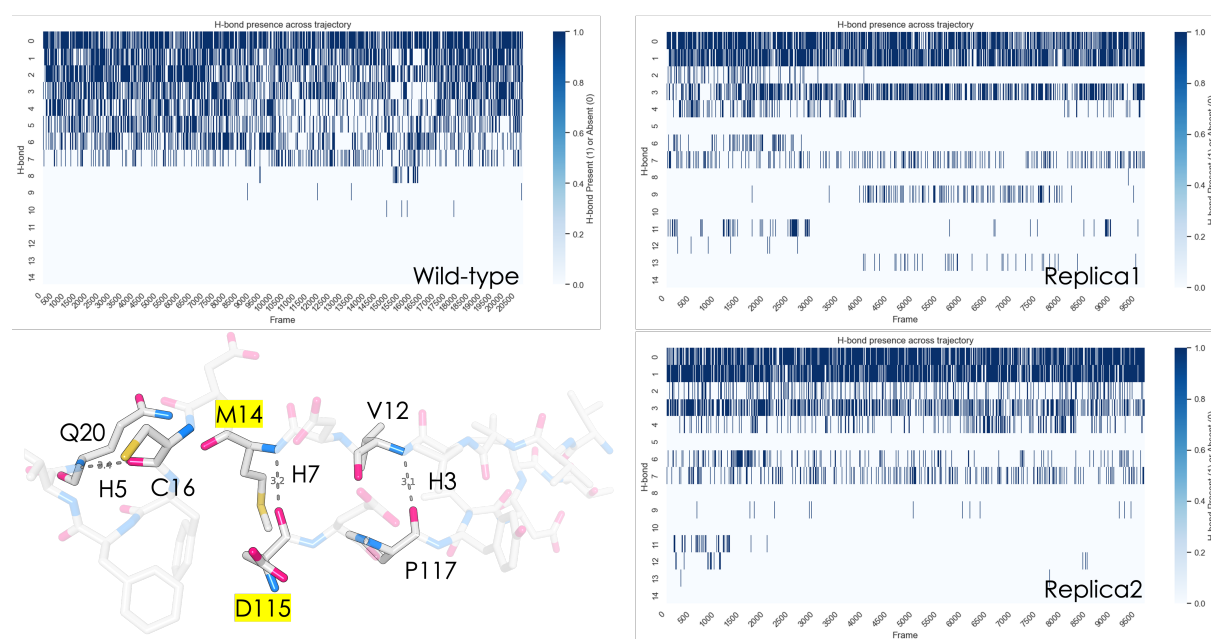


Figure C.1: **Rearrangement of active site hydrogen bonds in Ala117Pro.** Comparison of (upper left) the hydrogen-bond network involving residue 117 in the wild-type and (right) Ala117Pro mutant in two different replicas. (lower left) Structural view of the active site region highlighting the differences in hydrogen-bond interactions.

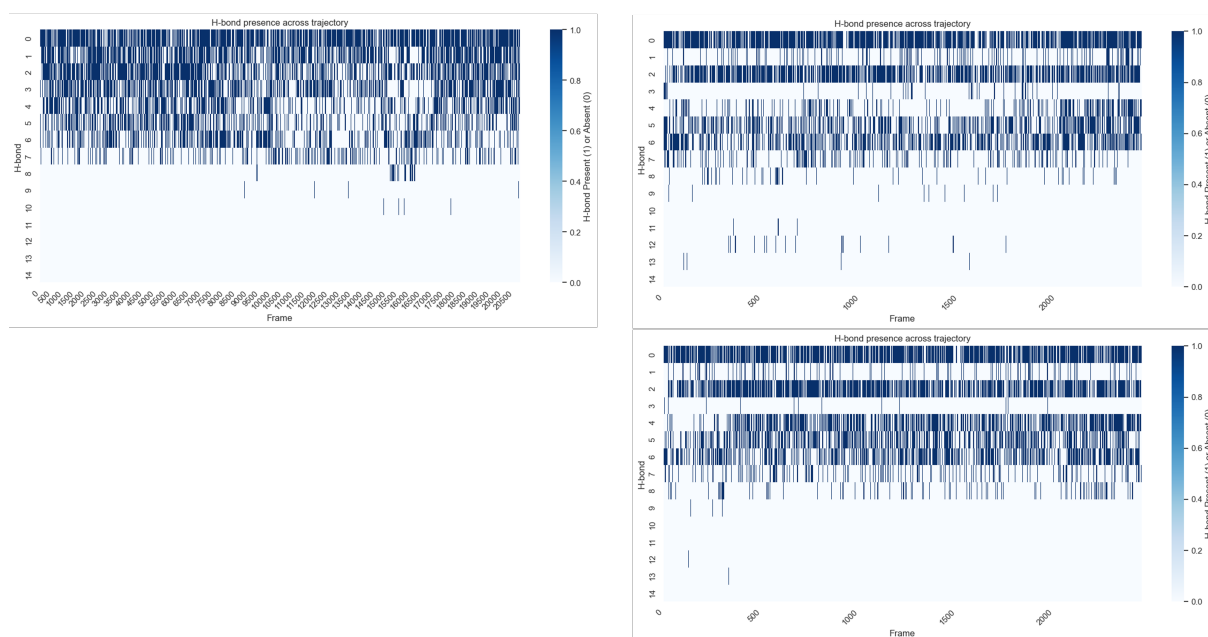


Figure C.2: **Rearrangement of active site hydrogen bonds in Thr122Pro.** Comparison of (upper left) the hydrogen-bond network involving residue 122 in the wild-type and (right) Thr122Pro mutant in two different replicas.

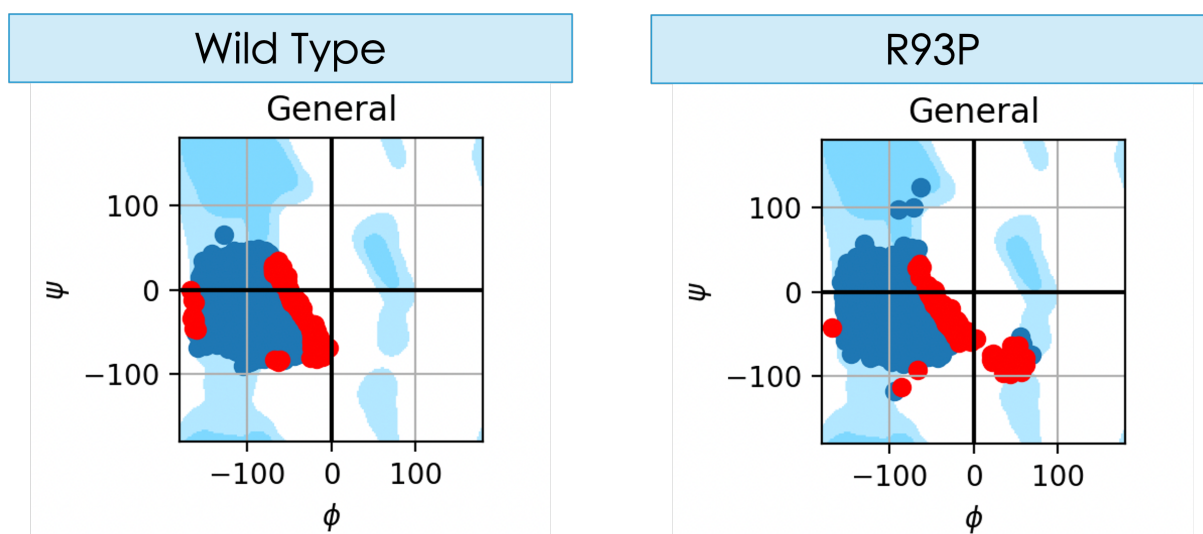


Figure C.3: **Changing in Ramachandran plot of Arg93Pro.** Comparison of the Ramachandran plot for (left) the wild-type and (right) Arg93Pro mutant, highlighting the differences in backbone dihedral angles.

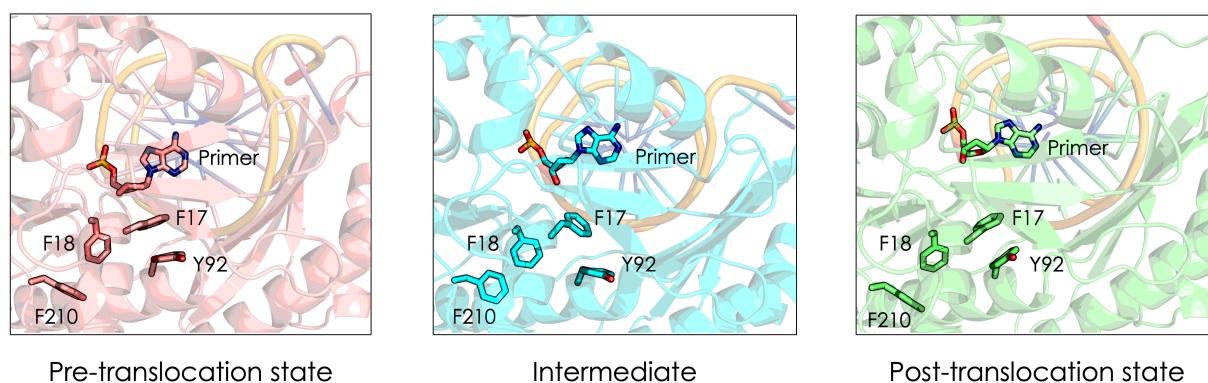


Figure C.4: **Steric gate network in wild-type Pol η .** Comparison of the steric gate network in (left) the wild-type pre-translocation state and (right) wild-type post-translocation state. In the middle panel, the structural rearrangement observed during translocation in the wild-type system is highlighted.

Bibliography

- [1] J. D. WATSON and F. H. C. CRICK. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 1953.
- [2] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Biochemistry*. Number 7. W. H. Freeman, New York, 2010.
- [3] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. Ucsf chimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82, 2021.
- [4] Rinku Jain, Aneel K. Aggarwal, and Olga Rechkoblit. Eukaryotic DNA polymerases. *Current Opinion in Structural Biology*, 53:77–87, 12 2018.
- [5] Wei Yang. An overview of Y-family DNA polymerases and a case study of human DNA polymerase η . *Biochemistry*, 53:2793–2803, 2014.
- [6] James M. Dewar and Johannes C. Walter. Mechanisms of DNA replication termination. *Nature Reviews Molecular Cell Biology*, 18(8):507–516, 2017.
- [7] Inacrist Geronimo, Pietro Vidossich, Elisa Donati, and Marco De Vivo. Computational investigations of polymerase enzymes: Structure, function, inhibition, and biotechnology. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11(6), 11 2021.
- [8] Elisa Donati, Pietro Vidossich, and Marco De Vivo. Molecular Mechanism of Phosphate Steering for DNA Binding, Cleavage Localization, and Substrate Release in Nucleases. *ACS Catalysis*, 11:13244–13254, 10 2021.
- [9] Mathias Girbig, Agata D. Misiaszek, and Christoph W. Müller. Structural insights into nuclear transcription by eukaryotic DNA-dependent RNA polymerases. *Nature Reviews Molecular Cell Biology*, 23(9):603–622, 2022.
- [10] Sabine S. Lange, Kei-ichi Takata, and Richard D. Wood. DNA polymerases and cancer. *Nature Reviews Cancer*, 11(2):96–110, 2011.
- [11] Lawrence A. Loeb and Raymond J. Monnat. DNA polymerases and human disease. *Nature Reviews Genetics*, 9:594–604, 2008.
- [12] Yazdan Maghsoud, Arkanil Roy, Emmett M. Leddin, and G. Andre´s Cisneros. Effects of the Y432S Cancer-Associated Variant on the Reaction Mechanism of Human DNA Polymerase κ . *Journal of Chemical Information and Modeling*, 64(10):4231–4249, 2024.
- [13] Chikahide Masutani, Rika Kusumoto, Ayumi Yamada, Naoshi Dohmae, Masayuki Yokoi, Mayumi Yuasa, Marito Araki, Shigenori Iwai, Koji Takio, and Fumio Hanaoka. The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase η . *Nature*, 399(6737):700–704, 1999.

- [14] Todd C. Appleby, Jason K. Perry, Eisuke Murakami, Ona Barauskas, Joy Feng, Aesop Cho, David Fox III, Diana R. Wetmore, Mary E. McGrath, Adrian S. Ray, Michael J. Sofia, S. Swaminathan, and Thomas E. Edwards. Structural basis for RNA replication by the hepatitis C virus polymerase. *Science*, 347(6223):771–775, 2015.
- [15] Joseph T Kim, Andrew D Hamilton, Christopher M Bailey, Robert A Domoal, Robert A Domoal, Ligong Wang, Karen S Anderson, and William L Jorgensen. FEP-Guided Selection of Bicyclic Heterocycles in Lead Optimization for Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Journal of the American Chemical Society*, 128(48):15372–15373, 2006.
- [16] Manli Wang, Ruiyuan Cao, Leike Zhang, Xinglou Yang, Jia Liu, Mingyue Xu, Zhengli Shi, Zhihong Hu, Wu Zhong, and Gengfu Xiao. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Research*, 30(3):269–271, 2020.
- [17] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 59(9):4035–4061, 2016.
- [18] Qi Peng, Ruchao Peng, Bin Yuan, Min Wang, Jingru Zhao, Lifeng Fu, Jianxun Qi, and Yi Shi. Structural Basis of SARS-CoV-2 Polymerase Inhibition by Favipiravir. *The Innovation*, 2(1):100080, 2021.
- [19] William Fried, Mrityunjay Tyagi, Leonid Minakhin, Gurushankar Chandramouly, Taylor Tredinick, Mercy Ramanjulu, William Auerbacher, Marissa Calbert, Timur Rusanov, Trung Hoang, Nikita Borisonnik, Robert Betsch, John J. Krais, Yifan Wang, Umeshkumar M. Vekariya, John Gordon, George Morton, Tatiana Kent, Tomasz Skorski, Neil Johnson, Wayne Childers, Xiaojiang S. Chen, and Richard T. Pomerantz. Discovery of a small-molecule inhibitor that traps Pol θ on DNA and synergizes with PARP inhibitors. *Nature Communications*, 15(1):2862, 2024.
- [20] Xiaowei Wang, Tiantian Xu, Yuan Yao, Peter Pak-Hang Cheung, Xin Gao, and Lu Zhang. SARS-CoV-2 RNA-Dependent RNA Polymerase Follows Asynchronous Translocation Pathway for Viral Transcription and Replication. *The Journal of Physical Chemistry Letters*, 14(45):10119–10128, 2023.
- [21] Adithya Nair and Zoltán Kis. Bacteriophage RNA polymerases: catalysts for mRNA vaccines and therapeutics. *Frontiers in Molecular Biosciences*, 11:1504876, 2024.
- [22] Federico Munafò, Michela Nigro, Nicoletta Brindani, Jacopo Manigrasso, Inacrist Geronimo, Giuliana Ottonello, Andrea Armirotti, and Marco De Vivo. Computer-aided identification, synthesis, and biological evaluation of DNA polymerase η inhibitors for the treatment of cancer. *European Journal of Medicinal Chemistry*, 248, 2 2023.
- [23] Kary B. Mullis and Fred A. Faloona. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. volume 155 of *Methods in Enzymology*, pages 335–350. 1987.
- [24] Russell Higuchi, Carita Fockler, Gavin Dollinger, and Robert Watson. Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions, 1993.

- [25] Inacrist Geronimo, Pietro Vidossich, and Marco De Vivo. Local Structural Dynamics at the Metal-Centered Catalytic Site of Polymerases is Critical for Fidelity. *ACS Catalysis*, 11:14110–14121, 11 2021.
- [26] Vito Genna, Elisa Donati, and Marco De Vivo. The Catalytic Mechanism of DNA and RNA Polymerases. *ACS Catalysis*, 8:11103–11118, 12 2018.
- [27] Alessia Visigalli, Enrico Trizio, Luigi Bonati, Pietro Vidossich, Michele Parrinello, and Marco De Vivo. Coordinated Residue Motions at the Enzyme–Substrate Interface Promote DNA Translocation in Polymerases. *Journal of the American Chemical Society*, 147(26):22972–22985, 2025.
- [28] David Ricardo Figueroa Blanco, Pietro Vidossich, and Marco De Vivo. Correct Nucleotide Selection Is Confined at the Binding Site of Polymerase Enzymes. *Journal of Chemical Information and Modeling*, 64(13):5285–5294, 2024.
- [29] Upeksha C. Dissanayake, Arkanil Roy, Yazdan Maghsoud, Sarthi Polara, Tanay Debnath, and G. Andrés Cisneros. Computational studies on the functional and structural impact of pathogenic mutations in enzymes. *Protein Science*, 34(4):e70081, 2025.
- [30] Lakindu S. Pathira Kankanamge, Alexandra Mora, Mary Jo Ondrechen, and Penny J. Beuning. Biochemical Activity of 17 Cancer-Associated Variants of DNA Polymerase Kappa Predicted by Electrostatic Properties. *Chemical Research in Toxicology*, 36(11):1789–1803, 2023.
- [31] Yih Wen Chen, James E. Cleaver, Fumio Hanaoka, Ching Fang Chang, and Kai Ming Chou. A novel role of DNA polymerase η in modulating cellular sensitivity to chemotherapeutic agents. *Molecular Cancer Research*, 4:257–265, 2006.
- [32] Olga Rechkoblit, Jayati Roy Choudhury, Angeliki Buku, Louise Prakash, Satya Prakash, and Aneel K. Aggarwal. Structural basis for polymerase η -promoted resistance to the anticancer nucleoside analog cytarabine. *Scientific Reports*, 8, 12 2018.
- [33] Christian Biertümpfel, Ye Zhao, Yuji Kondo, Santiago Ramón-Maiques, Mark Gregory, Jae Young Lee, Chikahide Masutani, Alan R. Lehmann, Fumio Hanaoka, and Wei Yang. Structure and mechanism of human DNA polymerase. *Nature*, 465:1044–1048, 6 2010.
- [34] Teruya Nakamura, Ye Zhao, Yuriko Yamagata, Yue Jin Hua, and Wei Yang. Watching DNA polymerase η make a phosphodiester bond. *Nature*, 487:196–201, 7 2012.
- [35] Vito Genna, Paolo Carloni, and Marco De Vivo. A Strategically Located Arg/Lys Residue Promotes Correct Base Paring During Nucleic Acid Biosynthesis in Polymerases. *Journal of the American Chemical Society*, 140(9):3312–3321, 3 2018.
- [36] Andrei A. Golosov, Joshua J. Warren, Lorena S. Beese, and Martin Karplus. The Mechanism of the Translocation Step in DNA Replication by DNA Polymerase I: A Computer Simulation Analysis. *Structure*, 18:83–93, 1 2010.

- [37] Daniel Adriano Silva, Dahlia R. Weiss, Fátima Pardo Avila, Lin Tai Da, Michael Levitt, Dong Wang, and Xuhui Huang. Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 111:7665–7670, 5 2014.
- [38] James R. Kiefer, Chen Mao, Jeffrey C. Braman, and Lorena S. Beese. Visualizing DNA replication in a catalytically active *Bacillus* DNA polymerase crystal. *Nature*, 391(6664):304–307, 1998.
- [39] François Boudsocq, Robert J. Kokoska, Brian S. Plosky, Alexandra Vaisman, Hong Ling, Thomas A. Kunkel, Wei Yang, and Roger Woodgate. Investigating the Role of the Little Finger Domain of Y-family DNA Polymerases in Low Fidelity Synthesis and Translesion Replication*. *Journal of Biological Chemistry*, 279(31):32932–32940, 2004.
- [40] Giacomo Janson and Alessandro Paiardini. PyMod 3: a complete suite for structural bioinformatics in PyMOL. *Bioinformatics*, 37(10):1471–1472, 2020.
- [41] Joonas A. Jamsen, David D. Shock, and Samuel H. Wilson. Watching right and wrong nucleotide insertion captures hidden polymerase fidelity checkpoints. *Nature Communications*, 13, 12 2022.
- [42] Vincent J. Cannistraro and John Stephen Taylor. DNA-Thumb Interactions and Processivity of T7 DNA Polymerase in Comparison to Yeast Polymerase η . *Journal of Biological Chemistry*, 279:18288–18295, 4 2004.
- [43] Wei Yang and Yang Gao. Translesion and Repair DNA Polymerases: Diverse Structure and Mechanism. 2018.
- [44] Jessica A. Brown and Zucui Suo. Unlocking the Sugar “Steric Gate” of DNA Polymerases. *Biochemistry*, 50(7):1135–1142, 2 2011.
- [45] Caleb Chang, Christie Lee Luo, Sarah Eleraky, Aaron Lin, Grace Zhou, and Yang Gao. Primer terminal ribonucleotide alters the active site dynamics of DNA polymerase η and reduces DNA synthesis fidelity. *Journal of Biological Chemistry*, 299, 3 2023.
- [46] Angela M. DeLucia, Nigel D.F. Grindley, and Catherine M. Joyce. An error-prone family Y DNA polymerase (DinB homolog from *Sulfolobus solfataricus*) uses a ‘steric gate’ residue for discrimination against ribonucleotides. *Nucleic Acids Research*, 31:4129–4137, 7 2003.
- [47] Katherine A. Donigan, Mary P. McLenigan, Wei Yang, Myron F. Goodman, and Roger Woodgate. The steric gate of dna polymerase ι regulates ribonucleotide incorporation and deoxyribonucleotide fidelity. *Journal of Biological Chemistry*, 289:9136–9145, 3 2014.
- [48] Katherine A. Donigan, Susana M. Cerritelli, John P. McDonald, Alexandra Vaisman, Robert J. Crouch, and Roger Woodgate. Unlocking the steric gate of DNA polymerase η leads to increased genomic instability in *Saccharomyces cerevisiae*. *DNA Repair*, 35:1–12, 11 2015.
- [49] Eitan Glick, Janice S. Chau, Kellie L. Vigna, Scott D. McCulloch, Elinor T. Adman, Thomas A. Kunkel, and Lawrence A. Loeb. Amino acid substitutions at conserved tyrosine 52 alter fidelity and

- bypass efficiency of human DNA polymerase η . *Journal of Biological Chemistry*, 278:19341–19346, 5 2003.
- [50] Justyna McIntyre. Polymerase iota - an odd sibling among Y family polymerases. *DNA Repair*, 86:102753, 2020.
- [51] Naoko Niimi, Akira Sassa, Atsushi Katafuchi, Petr Grúz, Hirofumi Fujimoto, Radha Rani Bonala, Francis Johnson, Toshihiro Ohta, and Takehiko Nohmi. The steric gate amino acid tyrosine 112 is required for efficient mismatched-primer extension by human DNA polymerase κ . *Biochemistry*, 48:4239–4246, 5 2009.
- [52] William A Beard and Samuel H Wilson. Structure and Mechanism of DNA Polymerase β .
- [53] M. Todd Washington, Robert E. Johnson, Satya Prakash, and Louise Prakash. Fidelity and Processivity of *Saccharomyces cerevisiae* DNA Polymerase η^* . *Journal of Biological Chemistry*, 274(52):36835–36838, 1999.
- [54] Eiji Ohashi, Katarzyna Bebenek, Toshiro Matsuda, William J. Feaver, Valerie L. Gerlach, Errol C. Friedberg, Haruo Ohmori, and Thomas A. Kunkel. Fidelity and Processivity of DNA Synthesis by DNA Polymerase κ , the Product of the Human DINB1 Gene*. *Journal of Biological Chemistry*, 275(50):39678–39684, 2000.
- [55] Kevin A. Fiala and Zucui Suo. Mechanism of DNA Polymerization Catalyzed by *Sulfolobus solfataricus* P2 DNA Polymerase IV †. *Biochemistry*, 43(7):2116–2125, 2004.
- [56] Katarzyna Bebenek, Toshiro Matsuda, Chikahide Masutani, Fumio Hanaoka, and Thomas A. Kunkel. Proofreading of DNA Polymerase η -dependent Replication Errors. *Journal of Biological Chemistry*, 276:2317–2320, 1 2001.
- [57] Agnès Tissier, John P. McDonald, Ekaterina G. Frank, and Roger Woodgate. *pol* ι , a remarkably error-prone human DNA polymerase. *Genes & Development*, 14(13):1642–1650, 2000.
- [58] Tyler M. Weaver, Luis M. Cortez, Thu H. Khoang, M. Todd Washington, Pratul K. Agarwal, and Bret D. Freudenthal. Visualizing Rev1 catalyze protein-template DNA synthesis. *Proceedings of the National Academy of Sciences*, 117(41):25494–25504, 2020.
- [59] Jing Wu, Alexandra De Paz, Bradley M. Zamft, Adam H. Marblestone, Edward S. Boyden, Konrad P. Kording, and Keith E.J. Tyo. DNA binding strength increases the processivity and activity of a Y-Family DNA polymerase. *Scientific Reports*, 7, 12 2017.
- [60] Tomoo Ogi, Patricia Kannouche, and Alan R. Lehmann. Localisation of human Y-family DNA polymerase κ : relationship to PCNA foci. *Journal of Cell Science*, 118(1):129–136, 2004.
- [61] I.R. Lehman, Maurice J. Bessman, Ernest S. Simms, and Arthur Kornberg. Enzymatic Synthesis of Deoxyribonucleic Acid I. PREPARATION OF SUBSTRATES AND PARTIAL PURIFICATION OF AN ENZYME FROM *ESCHERICHIA COLI*. *Journal of Biological Chemistry*, 233(1):163–170, 1958.

- [62] A. Lazcano, V. Llaca, G. E. Fox, and J. Oró. A classification of rna polymerases based on their evolutionary relatedness. *Origins of life and evolution of the biosphere*, 19(3-5):407–408, 1989.
- [63] Sherine S.L. Chan and William C. Copeland. DNA polymerase gamma and mitochondrial disease: Understanding the consequence of POLG mutations. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1787(5):312–319, 2009.
- [64] Laurie S. Kaguni. DNA POLYMERASE γ , THE MITOCHONDRIAL REPLICASE 1. *Biochemistry*, 73(1):293–320, 2004.
- [65] Mineaki Seki, Chikahide Masutani, Lee Wei Yang, Anthony Schuffert, Shigenori Iwai, Ivet Bahar, and Richard D Wood. High-efficiency bypass of DNA damage by human DNA polymerase η . *The EMBO Journal*, 23(22):4484–4494, 2004.
- [66] Chuxuan Li, Leora M. Maksoud, and Yang Gao. Structural basis of error-prone DNA synthesis by DNA polymerase θ . *Nature Communications*, 16(1):2063, 2025.
- [67] Tatiana Kent, Pedro A Mateos-Gomez, Agnel Sfeir, and Richard T Pomerantz. Polymerase θ is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *eLife*, 5:e13740, 2016.
- [68] Darius Kazlauskas, Mart Krupovic, Julien Guglielmini, Patrick Forterre, and Česlovas Venclovas. Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Research*, 48(18):gkaa760–, 2020.
- [69] Zuanning Yuan, Roxana Georgescu, Huilin Li, and Michael E. O’Donnell. Molecular choreography of primer synthesis by the eukaryotic pol α -primase. *Nature Communications*, 14(1):3697, 2023.
- [70] Erin E. Henninger and Zachary F. Pursell. DNA polymerase ϵ and its roles in genome stability. *IUBMB Life*, 66(5):339–351, 2014.
- [71] Claudia Lancey, Muhammad Tehseen, Vlad Stefan Raducanu, Fahad Rashid, Nekane Merino, Timothy J. Ragan, Christos G. Savva, Manal S. Zaher, Afnan Shirbini, Francisco J. Blanco, Samir M. Hamdan, and Alfredo De Biasio. Structure of the processive human Pol δ holoenzyme. *Nature Communications*, 11, 12 2020.
- [72] Young-Sam Lee, Mark T. Gregory, and Wei Yang. Human pol ζ purified with accessory subunits is active in translesion DNA synthesis and complements pol η in cisplatin bypass. *Proceedings of the National Academy of Sciences*, 111(8):2954–2959, 2014.
- [73] Jung-Hoon Yoon, Louise Prakash, and Satya Prakash. Error-free replicative bypass of (6–4) photo-products by DNA polymerase ζ in mouse and human cells. *Genes & Development*, 24(2):123–128, 2010.
- [74] José A. Morin, Francisco J. Cao, José M. Lázaro, J. Ricardo Arias-Gonzalez, José M. Valpuesta, José L. Carrascosa, Margarita Salas, and Borja Ibarra. Mechano-chemical kinetics of DNA replication: Identification of the translocation step of a replicative DNA polymerase. *Nucleic Acids Research*, 43:3643–3652, 2 2015.

- [75] Andrea J. Berman, Satwik Kamtekar, Jessica L. Goodman, José M. Lázaro, Miguel De Vega, Luis Blanco, Margarita Salas, and Thomas A. Steitz. Structures of phi29 DNA polymerase complexed with substrate: The mechanism of translocation in B-family polymerases. *EMBO Journal*, 26:3494–3505, 7 2007.
- [76] Linjing Yang, Karunesh Arora, William A. Beard, Samuel H. Wilson, and Tamar Schlick. Critical Role of Magnesium Ions in DNA Polymerase β 's Closing and Active Site Assembly. *Journal of the American Chemical Society*, 126(27):8441–8453, 2004.
- [77] Katarzyna Bebenek, Lars C. Pedersen, and Thomas A. Kunkel. Structure–Function Studies of DNA Polymerase λ . *Biochemistry*, 53(17):2781–2792, 2014.
- [78] Paula Andrade, María José Martín, Raquel Juárez, Francisco López de Saro, and Luis Blanco. Limited terminal transferase in human DNA polymerase μ defines the required balance between accuracy and efficiency in NHEJ. *Proceedings of the National Academy of Sciences*, 106(38):16203–16208, 2009.
- [79] Julian E. Sale, Alan R. Lehmann, and Roger Woodgate. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nature Reviews Molecular Cell Biology*, 13:141–152, 2012.
- [80] Scott D. McCulloch, Robert J. Kokoska, Chikahide Masutani, Shigenori Iwai, Fumio Hanaoka, and Thomas A. Kunkel. Preferential cis–syn thymine dimer bypass by DNA polymerase η occurs with biased fidelity. *Nature*, 428(6978):97–100, 2004.
- [81] Amritraj Patra, Leslie D Nagy, Qianqian Zhang, Yan Su, Livia Müller, F Peter Guengerich, and Martin Egli. Kinetics, Structure, and Mechanism of 8-Oxo-7,8-dihydro-2-deoxyguanosine Bypass by Human DNA Polymerase η . 2014.
- [82] Atsushi Katafuchi and Takehiko Nohmi. DNA polymerases involved in the incorporation of oxidized nucleotides into DNA: Their efficiency and template base preference. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*, 703:24–31, 11 2010.
- [83] Renee A. Beardslee, Samuel C. Suarez, Shannon M. Toffton, and Scott D. McCulloch. Mutation of the little finger domain in human DNA polymerase η alters fidelity when copying undamaged DNA. *Environmental and Molecular Mutagenesis*, 54(8):638–651, 2013.
- [84] Ye Zhao, Mark T. Gregory, Christian Biertümpfel, Yue Jin Hua, Fumio Hanaoka, and Wei Yang. Mechanism of somatic hypermutation at the WA motif by human DNA polymerase η . *Proceedings of the National Academy of Sciences of the United States of America*, 110:8146–8151, 2013.
- [85] Kristin A. Eckert. Nontraditional Roles of DNA Polymerase Eta Support Genome Duplication and Stability. *Genes*, 14:175, 1 2023.
- [86] Andrew Franklin, Peter J. Milburn, Robert V. Blanden, and Edward J. Steele. Human DNA polymerase- η , an A-T mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase. *Immunology and Cell Biology*, 82:219–225, 2004.

- [87] Sacha N Uljon, Robert E Johnson, Thomas A Edwards, Satya Prakash, Louise Prakash, and Aneel K Aggarwal. Crystal Structure of the Catalytic Core of Human DNA Polymerase Kappa. *Structure*, 12(8):1395–1404, 2004.
- [88] Kai-ming Chou. DNA Polymerase Eta and Chemotherapeutic Agents. *Antioxidants & Redox Signaling*, 14(12):2521–2529, 2011.
- [89] Timothy D. Silverstein, Robert E. Johnson, Rinku Jain, Louise Prakash, Satya Prakash, and Aneel K. Aggarwal. Structural basis for the suppression of skin cancers by DNA polymerase η . *Nature* 2010 465:7301, 465:1039–1043, 6 2010.
- [90] Bo Shu and Peng Gong. Structural basis of viral RNA-dependent RNA polymerase catalysis and translocation. *Proceedings of the National Academy of Sciences of the United States of America*, 113:E4005–E4014, 7 2016.
- [91] Wen Jin Wu, Wei Yang, and Ming Daw Tsai. How DNA polymerases catalyse replication and repair with contrasting fidelity. *Nature Reviews Chemistry*, 1:1–16, 8 2017.
- [92] Giulia Palermo, Andrea Cavalli, Michael L. Klein, Mercedes Alfonso-Prieto, Matteo Dal Peraro, and Marco De Vivo. Catalytic Metal Ions and Enzymatic Processing of DNA and RNA. *Accounts of Chemical Research*, 48(2):220–228, 2015.
- [93] Jacopo Manigrasso, Marco Marcia, and Marco De Vivo. Computer-aided design of RNA-targeted small molecules: A growing need in drug discovery. *Chem*, 7:2965–2988, 11 2021.
- [94] Elisa Donati, Vito Genna, and Marco De Vivo. Recruiting Mechanism and Functional Role of a Third Metal Ion in the Enzymatic Activity of 5' Structure-Specific Nucleases. *Journal of the American Chemical Society*, 142(6):2823–2834, 2020.
- [95] Vito Genna, Marco Marcia, and Marco De Vivo. A Transient and Flexible Cation- π Interaction Promotes Hydrolysis of Nucleic Acids in DNA and RNA Nucleases. *Journal of the American Chemical Society*, 141:10770–10776, 7 2019.
- [96] Vito Genna, Pietro Vidossich, Emiliano Ippoliti, Paolo Carloni, and Marco De Vivo. A Self-Activated Mechanism for Nucleic Acid Polymerization Catalyzed by DNA/RNA Polymerases. *Journal of the American Chemical Society*, 138:14592–14598, 11 2016.
- [97] Ming-Hsun Ho, Marco De Vivo, Matteo Dal Peraro, and Michael L. Klein. Understanding the Effect of Magnesium Ion Concentration on the Catalytic Activity of Ribonuclease H through Computation: Does a Third Metal Binding Site Modulate Endonuclease Catalysis? *Journal of the American Chemical Society*, 132(39):13702–13712, 2010.
- [98] Caleb Chang, Christie Lee Luo, and Yang Gao. In crystallo observation of three metal ion promoted DNA polymerase misincorporation. *Nature Communications*, 13(1):2346, 12 2022.
- [99] Eva Balint and Ildiko Unk. For the Better or for the Worse? The Effect of Manganese on the Activity of Eukaryotic DNA Polymerases. *International Journal of Molecular Sciences*, 25(1):363, 2023.

- [100] Eva Balint and Ildiko Unk. Selective Metal Ion Utilization Contributes to the Transformation of the Activity of Yeast Polymerase η from DNA Polymerization toward RNA Polymerization. *International Journal of Molecular Sciences*, 21(21):8248, 2020.
- [101] Ekaterina G. Frank and Roger Woodgate. Increased Catalytic Activity and Altered Fidelity of Human DNA Polymerase ϵ in the Presence of Manganese*. *Journal of Biological Chemistry*, 282(34):24689–24696, 2007.
- [102] Katie A. Wilson, Pedro A. Fernandes, Maria J. Ramos, and Stacey D. Wetmore. Exploring the Identity of the General Base for a DNA Polymerase Catalyzed Reaction Using QM/MM: The Case Study of Human Translesion Synthesis Polymerase η . *ACS Catalysis*, 9:2543–2551, 3 2019.
- [103] Jan Florián, Myron F. Goodman, and Arie Warshel. Computer Simulation of the Chemical Catalysis of DNA Polymerases: Discriminating between Alternative Nucleotide Insertion Mechanisms for T7 DNA Polymerase. *Journal of the American Chemical Society*, 125(27):8163–8177, 2003.
- [104] David R. Stevens and Sharon Hammes-Schiffer. Exploring the Role of the Third Active Site Metal Ion in DNA Polymerase η with QM/MM Free Energy Simulations. *Journal of the American Chemical Society*, 140(28):8965–8969, 2018.
- [105] Vito Genna, Roberto Gaspari, Matteo Dal Peraro, and Marco De Vivo. Cooperative motion of a key positively charged residue and metal ions for DNA replication catalyzed by human DNA Polymerase. *Nucleic Acids Research*, 44:2827–2836, 2016.
- [106] Vito Genna, Matteo Colombo, De Vivo, Marco, and Marco Marcia. Second-Shell Basic Residues Expand the Two-Metal-Ion Architecture of DNA and RNA Processing Enzymes. *Structure*, 26(1):40–50.e2, 1 2018.
- [107] Jacopo Manigrasso, Isabel Chillón, Vito Genna, Pietro Vidossich, Srinivas Somarowthu, Anna Marie Pyle, Marco De Vivo, and Marco Marcia. Visualizing group II intron dynamics between the first and second steps of splicing. *Nature communications*, 11, 12 2020.
- [108] Pietro Vidossich, Luis Eduardo Castañeda Moreno, Cristiano Mota, Daniele de Sanctis, Gian Pietro Miscione, and Marco De Vivo. Functional Implications of Second-Shell Basic Residues for dUTPase DR2231 Enzymatic Specificity. *ACS Catalysis*, 10(23):13825–13833, 2020.
- [109] Wonmuk Hwang and Martin Karplus. Structural basis for power stroke vs. Brownian ratchet mechanisms of motor proteins. *Proceedings of the National Academy of Sciences*, 116(40):19777–19785, 2019.
- [110] Lin-Tai Da, Chao E, Yao Shuai, Shaogui Wu, Xiao-Dong Su, and Jin Yu. T7 RNA polymerase translocation is facilitated by a helix opening on the fingers domain that may also prevent backtracking. *Nucleic Acids Research*, 45(13), 2017.
- [111] Sylvie Doublié, Stanley Tabor, Alexander M. Long, Charles C. Richardson, and Tom Ellenberger. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature* 1998 391:6664, 391:251–258, 1 1998.

- [112] Michael Feig and Zachary F. Burton. RNA Polymerase II with Open and Closed Trigger Loops: Active Site Dynamics and Nucleic Acid Translocation. *Biophysical Journal*, 99(8):2577–2586, 2010.
- [113] Olga Rechkoblit, Lucy Malinina, Yuan Cheng, Vitaly Kuryavyi, Suse Broyde, Nicholas E. Geacintov, and Dinshaw J. Patel. Stepwise translocation of Dpo4 polymerase during error-free bypass of an oxoG lesion. *PLoS Biology*, 4:0025–0042, 2006.
- [114] Ryan C. Wilson and Janice D. Pata. Structural Insights into the Generation of Single-Base Deletions by the Y Family DNA Polymerase Dbh. *Molecular Cell*, 29(6):767–779, 2008.
- [115] Xiaowei Wang, Yuan Yao, Xin Gao, and Lu Zhang. SARS-CoV-2 RdRp Follows Asynchronous Translocation Pathway for Viral Transcription and Replication. 2022.
- [116] Sefora Naomi Agro`, Riccardo Rozza, Santiago Movilla, Jana Aupicč, and Alessandra Magistrato. Molecular Dynamics Simulations Elucidate the Molecular Basis of Pre-mRNA Translocation by the Prp2 Spliceosomal Helicase. *Journal of Chemical Information and Modeling*, 63(13):4180–4189, 2023.
- [117] Adriaán García-Martínez, Kirill Zinovjev, José Javier Ruiz-Pernía, and Iñaki Tuñoón. Conformational Changes and ATP Hydrolysis in Zika Helicase: The Molecular Basis of a Biomolecular Motor Unveiled by Multiscale Simulations. *Journal of the American Chemical Society*, 145(45):24809–24819, 2023.
- [118] Lin-Tai Da, Dong Wang, and Xuhui Huang. Dynamics of Pyrophosphate Ion Release and Its Coupled Trigger Loop Motion from Closed to Open State in RNA Polymerase II. *Journal of the American Chemical Society*, 134(4):2399–2406, 2012.
- [119] Janice D. Pata. Structural diversity of the Y-family DNA polymerases. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(5):1124–1135, 2010.
- [120] Inacrist Geronimo, Pietro Vidossich, and Marco De Vivo. On the Role of Molecular Conformation of the 8-Oxoguanine Lesion in Damaged DNA Processing by Polymerases. *Journal of Chemical Information and Modeling*, 63:1521–1528, 3 2023.
- [121] Zhixiang Wu, Lianxin Peng, Yichen Hu, Tao Xie, Hailian Yan, Hua Wan, Wei Liu, Li Liang, Zhenjian Xie, Kunping Liu, Hubing Shi, Gang Zhao, and Jianping Hu. BP[dG]-induced distortions to DNA polymerase and DNA duplex: A detailed mechanism of BP adducts blocking replication. *Food and Chemical Toxicology*, 140:111325, 2020.
- [122] Ajay Ummat, Olga Rechkoblit, Rinku Jain, Jayati Roy Choudhury, Robert E. Johnson, Timothy D. Silverstein, Angeliki Buku, Samer Lone, Louise Prakash, Satya Prakash, and Aneel K. Aggarwal. Structural basis for cisplatin DNA damage tolerance by human polymerase η during cancer chemotherapy. *Nature Structural and Molecular Biology*, 19:628–632, 2012.
- [123] Huan Dong, Radha Rani Bonala, Naomi Suzuki, Francis Johnson, Arthur P. Grollman, and Shinya Shibutani. Mutagenic Potential of Benzo[a]pyrene-Derived DNA Adducts Positioned in Codon 273 of the Human P53 Gene †. *Biochemistry*, 43(50):15922–15928, 2004.

- [124] M. Todd Washington, Louise Prakash, and Satya Prakash. Mechanism of nucleotide incorporation opposite a thymine-thymine dimer by yeast DNA polymerase η . *Proceedings of the National Academy of Sciences of the United States of America*, 100:12093–12098, 10 2003.
- [125] Scott D. McCulloch, Robert J. Kokoska, Chikahide Masutani, Shigenori Iwai, Fumio Hanaoka, and Thomas A. Kunkel. Preferential cis-syn thymine dimer bypass by DNA polymerase η occurs with biased fidelity. *Nature*, 428(6978):97–100, 2004.
- [126] Elizaveta O. Boldinova, Artem Ignatov, Andrey Kulbachinskiy, and Alena V. Makarova. The active site residues Gln55 and Arg73 play a key role in DNA damage bypass by *S. cerevisiae* Pol η . *Scientific Reports* 2018 8:1, 8:1–11, 7 2018.
- [127] Robert E. Johnson, M. Todd Washington, Satya Prakash, and Louise Prakash. Fidelity of Human DNA Polymerase η . *Journal of Biological Chemistry*, 275:7447–7450, 3 2000.
- [128] Yulia Pustovalova, Mariana T. Q. Magalhães, Sanjay D’Souza, Alessandro A. Rizzo, George Korza, Graham C. Walker, and Dmitry M. Korzhnev. Interaction between the Rev1 C-Terminal Domain and the PolD3 Subunit of Pol ζ Suggests a Mechanism of Polymerase Exchange upon Rev1/Pol ζ -Dependent Translesion Synthesis. *Biochemistry*, 55(13):2043–2053, 2016.
- [129] Eitan Glick, Kellie L. Vigna, and Lawrence A. Loeb. Mutations in human DNA polymerase η motif II alter bypass of DNA lesions. *The EMBO Journal*, 20(24):7303–7312, 2001.
- [130] Yan Su, Amritraj Patra, Joel M. Harp, Martin Egli, and F. Peter Guengerich. Roles of residues Arg-61 and Gln-38 of human DNA polymerase η in bypass of deoxyguanosine and 7,8-dihydro-8-oxo-2'-deoxyguanosine. *Journal of Biological Chemistry*, 290:15921–15933, 6 2015.
- [131] Vamsi K. Gali, Eva Balint, Nataliia Serbyn, Orsolya Frittmann, Francoise Stutz, and Ildiko Unk. Translesion synthesis DNA polymerase η exhibits a specific RNA extension activity and a transcription-associated function. *Scientific Reports*, 7(1):13055, 12 2017.
- [132] Bernard C Broughton, Agnes Cordonnier, Wim J Kleijer, Nicolaas G J Jaspers, Heather Fawcett, Anja Raams, Victor H Garritsen, Anne Stary, Marie-Françoise Avril, François Boudsocq, Chikahide Masutani, Fumio Hanaoka, Robert P Fuchs, Alain Sarasin, and Alan R Lehmann. Molecular analysis of mutations in DNA polymerase in xeroderma pigmentosum-variant patients. Technical report.
- [133] Bruno César Feltes and Carlos Frederico Martins Menck. Current state of knowledge of human DNA polymerase eta protein structure and disease-causing mutations. *Mutation Research - Reviews in Mutation Research*, 790, 7 2022.
- [134] Vikash Jha and Hong Ling. Structural basis of accurate replication beyond a bulky major benzo[a]pyrene adduct by human DNA polymerase kappa. *DNA Repair*, 49:43–50, 2017.
- [135] Kevin N. Kirouac and Hong Ling. Unique active site promotes error-free replication opposite an 8-oxo-guanine lesion by human DNA polymerase iota. *Proceedings of the National Academy of Sciences*, 108(8):3210–3215, 2011.

- [136] Deepak T Nair, Robert E Johnson, Louise Prakash, Satya Prakash, and Aneel K Aggarwal. Hoogsteen base pair formation promotes synthesis opposite the 1,N6-ethenodeoxyadenosine lesion by human DNA polymerase ϵ . *Nature Structural & Molecular Biology*, 13(7):619–625, 2006.
- [137] Yanbin Zhang, Fenghua Yuan, Xiaohua Wu, John-Stephen Taylor, and Zhigang Wang. Response of human DNA polymerase ϵ to DNA lesions. *Nucleic Acids Research*, 29(4):928–935, 2001.
- [138] Deepak T. Nair, Robert E. Johnson, Louise Prakash, Satya Prakash, and Aneel K. Aggarwal. DNA Synthesis across an Abasic Lesion by Yeast Rev1 DNA Polymerase. *Journal of Molecular Biology*, 406(1):18–28, 2011.
- [139] Chuan Tian, Koushik Kasavajhala, Kellon A A Belfon, Lauren Raguette, He Huang, Angela N Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, 16(1):528–552, 2019.
- [140] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.
- [141] Karl T. Debiec, David S. Cerutti, Lewis R. Baker, Angela M. Gronenborn, David A. Case, and Lillian T. Chong. Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *Journal of Chemical Theory and Computation*, 12(8):3926–3947, 2016.
- [142] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, Guillem Portella, Federica Battistini, Josep Lluís Gelpí, Carlos González, Michele Vendruscolo, Charles A Laughton, Sarah A Harris, David A Case, and Modesto Orozco. Parmbsc1: a refined force field for DNA simulations. *Nature Methods*, 13(1):55–58, 2016.
- [143] Olivia Love, Rodrigo Galindo-Murillo, Marie Zgarbová, Jiří Šponer, Petr Jurečka, and Thomas E. Cheatham. Assessing the Current State of Amber Force Field Modifications for DNA2023 Edition. *Journal of Chemical Theory and Computation*, 19(13):4299–4307, 2023.
- [144] Marie Zgarbová, Jiří Šponer, and Petr Jurečka. Refinement of the Sugar Puckering Torsion Potential in the AMBER DNA Force Field. *Journal of Chemical Theory and Computation*, 21(2):833–846, 2025.
- [145] Marie Zgarbová, Michal Otyepka, Jiří Šponer, Arnošt Mlaďek, Pavel Banář, Thomas E. Cheatham, and Petr Jurečka. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *Journal of Chemical Theory and Computation*, 7(9):2886–2902, 2011.
- [146] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73, 2017.

- [147] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.
- [148] In Suk Joung and Thomas E. Cheatham. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *The Journal of Physical Chemistry B*, 112(30):9020–9041, 2008.
- [149] Pengfei Li, Lin Frank Song, and Kenneth M. Merz. Systematic Parameterization of Monovalent Ions Employing the Nonbonded Model. *Journal of Chemical Theory and Computation*, 11(4):1645–1657, 2015.
- [150] Arkajyoti Sengupta, Zhen Li, Lin Frank Song, Pengfei Li, and Kenneth M. Merz. Parameterization of Monovalent Ions for the OPC3, OPC, TIP3P-FB, and TIP4P-FB Water Models. *Journal of Chemical Information and Modeling*, 61(2):869–880, 2021.
- [151] Xibing He, Viet H. Man, Wei Yang, Tai-Sung Lee, and Junmei Wang. A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics*, 153(11):114502, 2020.
- [152] H J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.
- [153] Langevin and P. Sur la théorie du mouvement brownien. *Comptes Rendus de l’Académie des Sciences (Paris)*, 146:530–533, 1908.
- [154] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984.
- [155] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695–1697, 1985.
- [156] Hans C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.
- [157] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007.
- [158] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.
- [159] Johan Aqvist, Petra Wennerström, Martin Nervall, Sinisa Bjelic, and Bjørn O. Brandsdal. Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. *Chemical Physics Letters*, 384(4-6):288–294, 2004.
- [160] Pekka Mark and Lennart Nilsson. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.

- [161] Rahman and A. Correlations in the motion of atoms in liquid argon. *Physical Review*, 136:A405–A411, 1964.
- [162] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.
- [163] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.
- [164] Giovanni Bussi and Davide Branduardi. Free-Energy Calculations with Metadynamics: Theory and Practice. Technical report, 2015.
- [165] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, 1999.
- [166] Michele Invernizzi, Pablo M. Piaggi, and Michele Parrinello. Unified Approach to Enhanced Sampling. *Physical Review X*, 10:41034, 2020.
- [167] Michele Invernizzi and Michele Parrinello. Rethinking Metadynamics: From Bias Potentials to Probability Distributions. *Journal of Physical Chemistry Letters*, 11:2731–2736, 4 2020.
- [168] Michele Invernizzi and Michele Parrinello. Exploration vs Convergence Speed in Adaptive-Bias Enhanced Sampling. *Journal of Chemical Theory and Computation*, 18:3988–3996, 6 2022.
- [169] Luigi Bonati. Training collective variables for enhanced sampling via neural networks based discriminant analysis. *arXiv*, 2021.
- [170] Luigi Bonati, Valerio Rizzi, and Michele Parrinello. Data-Driven Collective Variables for Enhanced Sampling. *Journal of Physical Chemistry Letters*, 11:2998–3004, 4 2020.
- [171] Enrico Trizio and Michele Parrinello. From Enhanced Sampling to Reaction Profiles. *The Journal of Physical Chemistry Letters*, 12(35):8621–8626, 2021.
- [172] Luigi Bonati, Enrico Trizio, Andrea Rizzi, and Michele Parrinello. A unified framework for machine learning collective variables for enhanced sampling simulations: mlcolvar. *Journal of Chemical Physics*, 159, 7 2023.
- [173] Enrico Trizio, Andrea Rizzi, Pablo M Piaggi, Michele Invernizzi, and Luigi Bonati. Advanced simulations with PLUMED: OPES and Machine Learning Collective Variables. *arXiv*, 2024.
- [174] Enrico Trizio, Peilin Kang, and Michele Parrinello. Everything everywhere all at once, a probability-based enhanced sampling approach to rare events. *arXiv*, 2024.
- [175] Mujahid Khan, B. K. Hooda, Arpit Gaur, Vikram Singh, Yogesh Jindal, Hemender Tanwar, Sushma Sharma, Sonia Sheoran, Dinesh Kumar Vishwakarma, Mohammad Khalid, Ghadah Shukri Albakri, Maha Awjan Alreshidi, Jeong Ryeol Choi, and Krishna Kumar Yadav. Ensemble and optimization algorithm in support vector machines for classification of wheat genotypes. *Scientific Reports*, 14(1):22728, 2024.

- [176] Yuyang Wang and Roni Khardon. Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part I. *Lecture Notes in Computer Science*, pages 711–727, 2012.
- [177] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
- [178] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neuroinformatics*, 7:21, 2013.
- [179] Michael Greenacre, Patrick J. F. Groenen, Trevor Hastie, Alfonso Iodice D’Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.
- [180] Steffen Schultze and Helmut Grubmüller. Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. *Journal of Chemical Theory and Computation*, 17(9):5766–5776, 2021.
- [181] Anna C. Belkina, Christopher O. Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E. Snyder-Cappione. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications*, 10(1):5415, 2019.
- [182] Arnaldo Rapallo. Upper-Order TICA and Fractional Non-Markovian Process to Model Anomalous Dynamic Regimes. *Journal of Chemical Theory and Computation*, 21(14):6711–6728, 2025.
- [183] Luigi Bonati, GiovanniMaria Piccini, and Michele Parrinello. Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences*, 118:e2113533118, 11 2021.
- [184] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.
- [185] Maroof K. Zafar, Leena Maddukuri, Amit Ketkar, Narsimha R. Pentthala, Megan R. Reed, Sarah Eddy, Peter A. Crooks, and Robert L. Eoff. A Small-Molecule Inhibitor of Human DNA Polymerase η Potentiates the Effects of Cisplatin in Tumor Cells. *Biochemistry*, 57(7):1262–1273, 2018.
- [186] Meihua Wang, Rui Li, Bo Shu, Xuping Jing, Han-Qing Ye, and Peng Gong. Stringent control of the RNA-dependent RNA polymerase translocation revealed by multiple intermediate structures. *Nature Communications*, 11(1):2605, 2020.
- [187] Thomas W. Kirby, Eugene F. DeRose, Nisha A. Cavanaugh, William A. Beard, David D. Shock, Geoffrey A. Mueller, Samuel H. Wilson, and Robert E. London. Metal-induced DNA translocation leads to DNA polymerase conformational activation. *Nucleic Acids Research*, 40(7):2974–2983, 2012.
- [188] M. Todd Washington, Robert E. Johnson, Satya Prakash, and Louise Prakash. Accuracy of thymine–thymine dimer bypass by *Saccharomyces cerevisiae* DNA polymerase η . *Proceedings of the National Academy of Sciences*, 97(7):3094–3099, 2000.

- [189] Mark Hedglin, Binod Pandey, and Stephen J. Benkovic. Characterization of human translesion DNA synthesis across a UV-induced DNA lesion. *eLife*, 5, 10 2016.
- [190] Alexander Knips and Martin Zacharias. Influence of a cis,syn-cyclobutane pyrimidine dimer damage on DNA conformation studied by molecular dynamics simulations. *Biopolymers*, 103(4):215–222, 2015.
- [191] Chikahide Masutani, Rika Kusumoto, Shigenori Iwai, and Fumio Hanaoka. Mechanisms of accurate translesion synthesis by human DNA polymerase η . *The EMBO Journal*, 19(12):3100–3109, 2000.
- [192] Bernard C. Broughton, Agnes Cordonnier, Wim J. Kleijer, Nicolaas G. J. Jaspers, Heather Fawcett, Anja Raams, Victor H. Garritsen, Anne Stary, Marie-Françoise Avril, François Boudsocq, Chikahide Masutani, Fumio Hanaoka, Robert P. Fuchs, Alain Sarasin, and Alan R. Lehmann. Molecular analysis of mutations in DNA polymerase η in xeroderma pigmentosum-variant patients. *Proceedings of the National Academy of Sciences*, 99(2):815–820, 2002.
- [193] Ye Zhao, Christian Biertümpfel, Mark T. Gregory, Yue Jin Hua, Fumio Hanaoka, and Wei Yang. Structural basis of human DNA polymerase η -mediated chemoresistance to cisplatin. *Proceedings of the National Academy of Sciences of the United States of America*, 109:7269–7274, 2012.
- [194] Atsushi Katafuchi, Akira Sassa, Naoko Niimi, Petr Grú, Hirofumi Fujimoto, Chikahide Masutani, Fumio Hanaoka, Toshihiro Ohta, and Takehiko Nohmi. Critical amino acids in human DNA polymerases γ and δ involved in erroneous incorporation of oxidized nucleotides.
- [195] Philip Nevin, John R. Engen, and Penny J. Beuning. Steric gate residues of Y-family DNA polymerases DinB and pol kappa are crucial for dNTP-induced conformational change. *DNA Repair*, 29:65–73, 5 2015.
- [196] Enrico Trizio, Peilin Kang, and Michele Parrinello. Everything everywhere all at once, a probability-based enhanced sampling approach to rare events. *arXiv*, 2024.
- [197] Katie A. Wilson, Pedro A. Fernandes, Maria J. Ramos, and Stacey D. Wetmore. Exploring the Identity of the General Base for a DNA Polymerase Catalyzed Reaction Using QM/MM: The Case Study of Human Translesion Synthesis Polymerase η . *ACS Catalysis*, 9(3):2543–2551, 2019.
- [198] D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Ko valenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O’Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, and Kollman, and P.A. Amber 2022 Reference Manual. 2022.

- [199] Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, 2013.
- [200] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.
- [201] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.
- [202] Andrej Sali and Tom L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993.
- [203] David A. Case, Hasan Metin Aktulga, Kellon Belfon, David S. Cerutti, G. Andrés Cisneros, Vinícius Wilian D. Cruzeiro, Negin Forouzesh, Timothy J. Giese, Andreas W. Götz, Holger Gohlke, Saeed Izadi, Koushik Kasavajhala, Mehmet C. Kaymak, Edward King, Tom Kurtzman, Tai-Sung Lee, Pengfei Li, Jian Liu, Tyler Luchko, Ray Luo, Madushanka Manathunga, Matias R. Machado, Hai Minh Nguyen, Kurt A. O’Hearn, Alexey V. Onufriev, Feng Pan, Sergio Pantano, Ruxi Qi, Ali Rahnamoun, Ali Risheh, Stephan Schott-Verdugo, Akhil Shajan, Jason Swails, Junmei Wang, Haixin Wei, Xiongwu Wu, Yongxian Wu, Shi Zhang, Shiji Zhao, Qiang Zhu, Thomas E. Cheatham, Daniel R. Roe, Adrian Roitberg, Carlos Simmerling, Darrin M. York, Maria C. Nagan, and Kenneth M. Merz. AmberTools. *Journal of Chemical Information and Modeling*, 63(20):6183–6191, 2023.
- [204] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *arXiv*, 2013.
- [205] D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O’Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and Kollman, and P.A. Amber 2021. 2021.
- [206] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2):283–291, 1993.
- [207] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.