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

Computational Approaches for Studying Enzyme Function and Protein-Ligand Interactions: From atomistic simulations to deep learning

Presentata da: David Ricardo Figueroa Blanco

Coordinatore Dottorato

Daniele Bonacorsi

Supervisore

Marco de Vivo

Co-supervisore

Andrea Cavallli

Esame finale anno 2026

DEDICATION

A mi familia y especialmente a mis abuelas Melida y Marina.

ACKNOWLEDGEMENTS

I want to sincerely thank my supervisor, Marco De Vivo. I'm deeply grateful for his trust and for an opportunity that truly changed my perspective on science. His guidance on writing an impactful paper was invaluable, and he showed me a vision of science that goes far beyond the editorial desk.

I am also incredibly grateful to Dr. Pietro Vidossich. His constant mentorship, patience, and insightful scientific discussions were essential to my doctoral studies. He has been a true inspiration, both as a rigorous scientist and a supportive colleague.

A huge thanks goes to all the members of my lab at the Istituto Italiano di Tecnologia (IIT). The collaborative and stimulating environment was crucial for my research and personal growth. I have to give a special thanks to Dr. Luigi Zanovello for his close and fruitful collaboration on designing and implementing the methods in Chapter 5.

I was also fortunate to collaborate with researchers outside my immediate group. My thanks to Dr. Gianni de Fabritiis, who I worked with on the project in Chapter 4 during my research stay abroad. I also really appreciate the insightful discussions with Prof. Massimiliano Pontil from the Computational Statistics and Machine Learning group at IIT, which were a huge benefit to my work. I was so lucky to share this journey with wonderful colleagues and friends in Genoa.

My sincere thanks to Joe, Ziggy, Manuel, Uma, Maria, Gian, Jacopo, and Elisa. Your support, friendship, and the cultural exchange we shared made this journey not only possible but truly wonderful.Finally, my deepest, heartfelt thanks go to my family—my parents, Ricardo and Carolina, and my sister, Camila—for their unconditional love and support, which has always been my foundation. To Anna, thank you for your incredible partnership, love, and for sharing this adventure with me.

TABLE OF CONTENTS

ABSTRACT

DNA and RNA polymerases are essential enzymes involved in diverse biological processes. Consequently, a detailed understanding of the molecular mechanisms underlying DNA and RNA synthesis is vital for elucidating the progression of diseases such as cancer. While experimental techniques like X-ray crystallography and steady-state kinetics have provided significant insights into replicative and DNA repair polymerases, computational methods offer atomic-level precision and a dynamic understanding of these processes. Furthermore, these methods facilitate the quantification of ligand interactions, often presenting advantages in terms of cost and the level of detail achievable compared to purely experimental approaches. This thesis explores the application of advanced computational methods such as molecular dynamics simulations, free energy calculations, and convolutional neural networks to investigate polymerase function and inhibition.

First, we investigate how mutations distant from the active site influence nucleotide selection in DNA polymerase β (Pol β). Through molecular dynamics simulations and free energy calculations, we demonstrate that the R283A mutation, located 14.8 Å from the catalytic site, reduces fidelity by disrupting critical interactions that stabilize correct base pairing geometries. Structural alignment reveals this mechanism is conserved across the polymerase X family. Second, we systematically benchmark alchemical free energy calculations for ribonucleotide discrimination in Pol β and RB69 polymerase. We demonstrate that while specialized nucleotide force fields like SHAW substantially outperform general-purpose parametrizations (GAFF2), significant errors persist for challenging nucleotide analogs. These limitations motivate the development of Neural Network Potential/Molecular Mechanics hybrid methods, for which we have curated a specialized training dataset. Finally, we develop a convolutional neural network for fragment-based drug design. Through chemistry-informed data curation and systematic optimization, our model showed reasonable accuracy across 30 fragment classes while demonstrating robust generalization to novel protein families. Validation on clinically relevant targets reveals both current capabilities and limitations that guide future development. Together,

these investigations advance our mechanistic understanding of polymerase fidelity and establish more reliable computational methods for predicting ligand binding and guiding inhibitor design.

PREFACE

This dissertation explores advanced computational methods for understanding DNA polymerases, which are enzymes fundamental to life and important therapeutic targets. The work sits at the intersection of structural biology, computational chemistry, and machine learning, addressing how molecular dynamics simulations, free energy calculations and convolutional neuronal network models can elucidate mechanism for fidelity, ribonucleotide discrimination and contribute to fragment based drug design.

The dissertation progresses from foundational concepts through methodological developments to practical applications:

Chapter 1 introduces polymerases as subjects of mechanistic study and targets for therapeutic intervention, reviewing current understanding of their function and computational approaches while identifying key knowledge gaps.

Chapter 2 provides the theoretical framework, covering molecular dynamics simulations, force fields, neural network potentials, alchemical free energy calculations, and deep learning architectures employed throughout this work.

Chapter 3 investigates nucleotide selection in DNA polymerase β, revealing how a distant arginine residue critically influences fidelity by stabilizing productive geometries only for correct base pairs through molecular dynamics and free energy calculations.

Chapter 4 evaluates the Alchemical Transfer Method for predicting ribonucleotide discrimination, comparing force fields and demonstrating that while specialized parameterizations improve predictions, achieving reliable results for diverse nucleotide analogs requires neural network potentials beyond classical approaches.

Chapter 5 develops a convolutional neural network for fragment-based drug design, framing molecular generation as classification to ensure chemical validity. Through systematic optimization and validation on clinical targets, this work demonstrates both promise and current limitations of deep learning for structure-based design.

Chapter 6 synthesizes insights across chapters, discussing broader implications and future directions.

1. INTRODUCTION

1.1. Polymerases: Central enzymes in DNA processing.

DNA and RNA polymerases are enzymes that catalyze the synthesis of nucleic acids, which is crucial for the storage of genetic information in all biological systems.[1,2] These enzymes are central to life itself, executing critical processes such as DNA replication, DNA repair, recombination, and transcription, thereby ensuring cell growth, the propagation of life, and the maintenance of genome stability.[2,3] Cells use a diverse array of DNA polymerases, categorized into distinct families (e.g., A, B, X, Y), each with specialized functions in various aspects of DNA processes.[2] Given their fundamental roles, the dysfunction and dysregulation are often implicated in the pathogenesis of cancer and in the propagation of viral and bacterial infections.[4,5]

A profound understanding of the catalytic mechanisms, structural dynamics, substrate specificity, and accuracy of polymerases is essential to decipher how these enzymes accomplish their diverse and intricate tasks, including how they respond to challenges such as DNA damage or the presence of nucleotide analogs.[6,7] These mechanistic insights, which include delineating complex conformational changes, the roles of metal ions in catalysis, and the principles of substrate selection, are the basis for comprehending their biological functions and for devising strategies to modulate their activity.[8–10] Consequently, polymerases have emerged as highly significant targets for therapeutic intervention in a range of human diseases. The development of inhibitors that selectively target viral or bacterial polymerases, or that exploit the dependencies of cancer cells

on specific DNA replication and repair polymerases, opens a new direction in treating different diseases.[11] Beyond their biological roles, the catalytic capabilities of polymerases have been harnessed as indispensable tools in modern molecular biology and biotechnology, powering transformative techniques such as the polymerase chain reaction (PCR), DNA sequencing, and various applications in synthetic biology.[1,12] In this context, computational methodologies are playing an increasingly critical role by providing atomic-level resolution of polymerase action and interactions with inhibitors, thereby complementing experimental investigations, guiding new therapeutic strategies, and exploring relevant biotechnological applications.

1.2. Polymerases: Function, Properties and Classification

DNA and RNA polymerases build new nucleic acid strands by adding nucleotides (dNTPs for DNA, rNTPs for RNA) to the 3' end of a primer chain, using an existing strand as a template. Two main features measure the efficiency of this process: fidelity and processivity. Fidelity is the polymerase's accuracy in choosing the correct nucleotide that pairs with the template (A with T/U, and G with C). This accuracy can be measured by comparing how efficiently of the enzyme incorporates incorrect versus correct nucleotides, specifically in pre-steady state kinetics, this is measured by the ratio $(k_{cat}/K_M)_{incorrect}/(k_{cat}/K_M)_{correct}$. Processivity refers to the number of nucleotides a polymerase adds in one binding event before detaching from the template. This can be determined by experiments that track nucleotide addition during a single binding event or by calculating the ratio of the nucleotide incorporation rate (k_{pol}) to the enzyme's dissociation rate (k_{off}) from the primer template. Both fidelity and processivity vary considerably among polymerases, suiting their specific jobs in a cell, such as copying DNA or repairing it.[1,13]

Polymerases are phylogenetically and structurally classified into seven major families—A, B, C, D, X, Y, and reverse transcriptases (RTs)—based on conserved sequence motifs and

three-dimensional fold features. Family A enzymes (e.g., bacterial Pol I, mitochondrial Pol γ) often possess both 3′→5′ exonuclease and 5′→3′ exonuclease activities that facilitate the removal of mismatches, serving in replication and repair tasks.[14] Family B polymerases (e.g., eukaryotic Pol δ/ε, bacterial Pol II, bacteriophage T4/RB69 gp43) are the principal enzymes that duplicate genomes with high fidelity, aided by separate exonuclease domains too.[15] Family C comprises the multi-subunit bacterial Pol III holoenzyme, the main replicative polymerase in prokaryotes. Family D, found in archaea, represents a distinct replicative polymerase class with unique domain organization and proofreading modules. Family X polymerases (e.g., Pol β, λ, μ, TdT) are small, gap-filling enzymes active in base-excision repair and non-homologous end-joining, which are different processes involved in DNA repair, generally lacking intrinsic 3′→5′ proofreading.[16,17] Family Y enzymes (e.g., Pol η, κ, ι, Rev1) specialize in translesion synthesis, bypassing DNA lesions at the expense of lower fidelity and reduced processivity. Finally, reverse transcriptases (viral RTs and telomerase) perform RNA-templated DNA synthesis, critical for retroviral replication and telomere maintenance.[18,19]

Together, these families orchestrate all DNA replication and repair processes, exhibiting a broad spectrum of fidelities (error rates from ~10^{-1} to ~10^{-8}) and processivities (from distributive gap-fillers to highly processive replicases) aligned with their physiological roles in genome maintenance, damage tolerance, and information flow between DNA and RNA.

1.3. Common Structural Framework

Despite the wide range of polymerase activities, all DNA polymerases share a conserved “right-hand” architecture in their catalytic domain, with three interlocking subdomains: C (Catalytic), N (nascent base-pair binding), and D (DNA Binding), that together coordinate substrate binding and catalysis (Figure 1A). The catalytic subdomain contains invariant acidic

residues that coordinate two catalytic divalent metal ions (Mg^{2+}), one metal ion is proposed to activate the 3'-hydroxyl group of the primer for nucleophilic attack on the α-phosphate of the incoming deoxynucleoside triphosphate (dNTP), while the second metal ion coordinates the dNTP, stabilizes the negative charge of the transition state, and facilitates the departure of pyrophosphate.[17] Conserved catalytic motifs also interact directly with the triphosphate of the incoming dNTP, ensuring precise alignment for in-line nucleophilic attack on the α‑phosphate (Figure 1B).[1,7,8,17] The N domain is a more dynamic region that interacts with the incoming dNTP and the template strand. Upon binding of a correctly base-paired dNTP, the alpha helix undergoes a significant conformational change, transitioning from an "open" to a "closed" state. This movement, which often involves conserved α-helical, helps to precisely orient the dNTP within the active site for catalysis.[20–22] The D domain (often called palm domain) completes the characteristic right-hand architecture and primarily interacts with the DNA duplex, specifically engaging the primer-template junction and the newly synthesized double-stranded DNA.[23] This interaction is vital for anchoring the polymerase to its DNA substrate, therefore enhancing the stability of the polymerase-DNA complex, correctly positioning the DNA within the catalytic cleft, and promoting processivity.

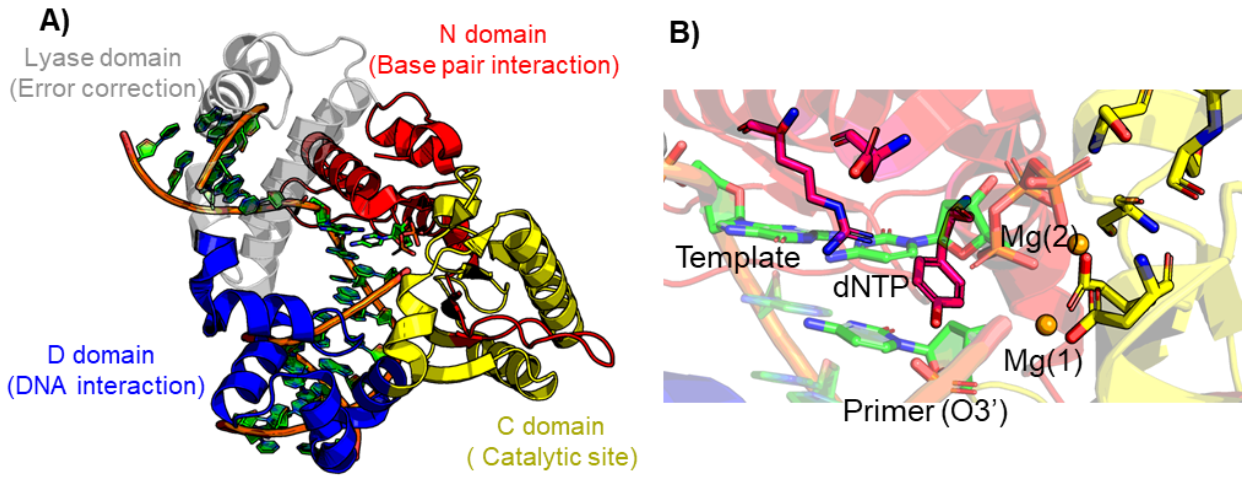


Figure 1. A) The characteristic "right-hand" structure showing three core subdomains: C domain (catalytic site, yellow), N domain (nascent base-pair binding, red), and D domain (DNA binding, blue). DNA strands (template in orange, primer in green) are shown with the incoming dNTP at the active site. B) Detailed view of the catalytic site showing the two-metal-ion mechanism. Mg(1) activates the primer 3'-OH for nucleophilic attack while Mg(2) stabilizes the dNTP triphosphate.

While the catalytic core of DNA polymerases remains somehow conserved, many polymerases integrate additional specialized domains to ensure genomic integrity. In many high-fidelity polymerases, a separate 3′→5′ exonuclease domain—positioned between an N-terminal domain and the palm—proofreads errors by transferring mispaired primer ends from the polymerase site to the exonuclease site for excision, thereby reducing overall error rates by up to two orders of magnitude.[24,25] In contrast, some repair polymerases, like DNA polymerase β (Pol β) from the X-family, lack this function but possess other specialized modules, such as a 5'-deoxyribose phosphate (dRP) lyase domain essential for base excision repair (BER). Together, the catalytic, DNA-binding, and proofreading domains underlie the remarkable speed, selectivity, and versatility of DNA polymerases across all families and kingdoms.[17,26]

1.4. Catalytic Cycle

Polymerase action proceeds through a defined sequence of kinetic and structural steps that together balance rapid nucleotide addition with high fidelity (see Figure 2A) . First, the enzyme associates with the primer–template duplex in an “open” conformation (characterized by the N subdomain positioned away from the active site) to form the binary complex (see (I) in Figure 2A). Next, an incoming deoxynucleoside or ribonucleoside triphosphate forms a Watson–Crick pairing with the templating base; this nascent ternary complex is stabilized by contacts between the phosphate groups and the N domain as well as coordination to two divalent metal ions in the C domain.[10,17]

Upon correct base pairing, the N domain rotates inward by roughly 30–40°, transitioning the enzyme into a “closed” conformation that aligns the primer 3′-OH, the dNTP α-phosphate, and the catalytic carboxylates in the C domain, thereby enforcing an induced-fit checkpoint that discriminates against mismatches (Figure 2B). In this closed state, nucleotidyl transfer is catalyzed via a two-metal-ion mechanism in which Mg(1) activates the primer 3′-OH for attack on the α-phosphate and Mg(2) stabilizes the triphosphate and leaving pyrophosphate, resulting in phosphodiester bond formation and pyrophosphate release (Figure 2C). Extensive pre-steady-state kinetic analyses and high-resolution structural snapshots across multiple polymerase families confirm that these sequential events—substrate binding, induced fit, chemistry, and product release—underlie the exceptional speed and accuracy of DNA and RNA synthesis.[9,10,27–29]

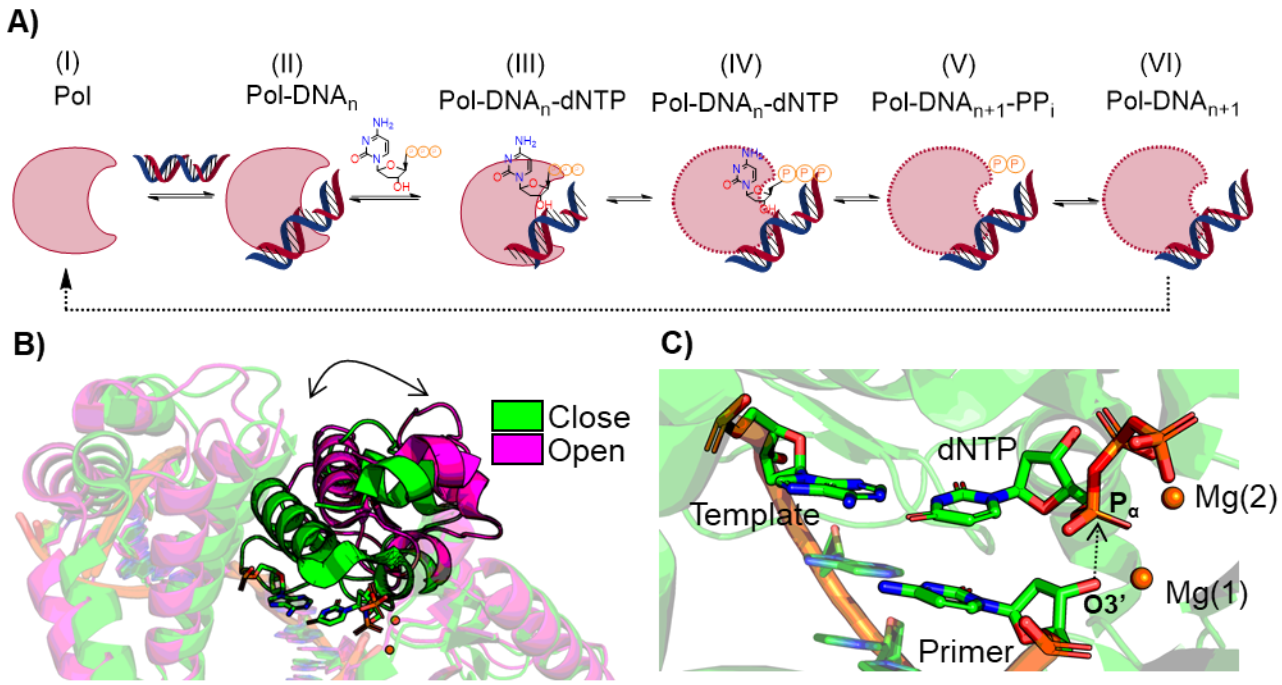


Figure 2. A) Schematic representation of the nucleotide incorporation cycle: (I) polymerase binds DNA, (II) incoming dNTP binds in the open conformation, (III) formation of the ternary complex with correct base pairing, (IV) conformational change to the closed state aligns catalytic residues, (V) phosphodiester bond formation and pyrophosphate (PPi) release, (VI) translocation to the next templating position. B) Structural overlay showing the open (pink) to closed (green) conformational transition of the N domain upon correct dNTP binding. C) Active site geometry in the closed state showing optimal alignment of the primer 3'-OH, incoming dNTP α-phosphate (Pα), and catalytic magnesium ions Mg(1) and Mg(2) for in-line nucleophilic attack and bond formation.

1.5. B-family polymerases (e.g., RB69) in high-fidelity DNA replication.

B-family DNA polymerases are the core replicases in many organisms (bacteriophages, archaea, eukaryotes). RB69 polymerase (gp43) from phage RB69 has long been a model for this family.[30] It is homologous in sequence and domain organization to eukaryotic replicative Pol α, δ, and ε Like other B-family enzymes, RB69pol is a "right-hand" with an N-terminal 3′–5′ exonuclease domain and a C-terminal polymerase domain. This architecture forms a central cleft that binds the primer‑template DNA. RB69pol's proofreading exonuclease gives overall error rates in the

10^{-6}–10^{-9} range.[31] Such high fidelity is critical for genome maintenance; indeed, perturbing base‑selective interactions in B-family polymerases rapidly raises mutagenesis.[31]

High-fidelity B-family polymerases use multiple checkpoints to ensure correct dNTP insertion. First, base recognition occurs in a nascent base‑pair binding pocket formed by conserved residues near the active site. These residues enforce correct Watson-Crick geometry through steric and chemical complementarity. Then the other mechanism for nucleotide selection starts. For example, the "fingers" domain optimally aligns substrates for catalysis, whereas an incorrect (mismatched) dNTP yields an unstable closed complex that reopens faster than chemistry can occur. Structural studies of RB69pol confirm these features.[32] Finally, interactions around the nascent base pair and the DNA minor groove also complements the binding of the incoming nucleotide. In the high-resolution ternary complex (PDB 3NCI), a network of four ordered water molecules occupies the DNA minor groove, hydrogen-bonding to N3/O2 of the template base. Upon formation of the closed complex, two of these waters are expelled, and multiple side-chain contacts (e.g., Gly568–Gly568-template N3, Ser565–template base face, O5′–C5 of the incoming dNTP) are optimized to stabilize the nascent base pair (Figure 3). These tightly packed polar and hydrophobic interactions enhance Watson–Crick selectivity for correct bases. In addition, minor-groove hydrogen bonds from polymerase or water at the primer/template positions n–1 and n–2 help stabilize correct nascent pairs, while mismatches create gaps or steric clashes that further destabilize misinsertion. Altogether, RB69pol (and other B-family replicases) require a well-aligned nascent base pair and proper metal coordination to catalyze extension, so that incorrect dNTPs are incorporated orders of magnitude more slowly.[30]

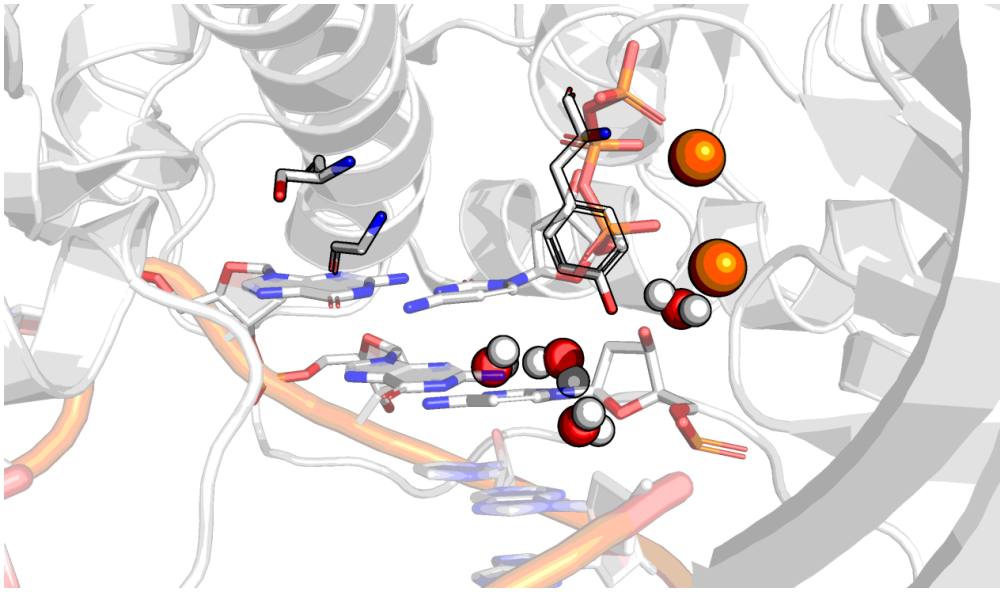


Figure 3. Detailed view of the catalytic site with the template-dNTP base pair positioned for Pol RB69 catalysis. Four ordered water molecules (red and white spheres) occupy the DNA minor groove, forming hydrogen bonds with the N3/O2 atoms of the template base. Two catalytic magnesium ions (orange spheres) coordinate the primer 3'-OH and dNTP triphosphate. Upon closure of the N domain, two of these waters are expelled while specific side-chain interactions stabilize Watson-Crick geometry, contributing to nucleotide selectivity.

1.6. Family X (Pol β) involvement in base-excision repair.

DNA polymerase β (Pol β) is the prototypical X-family DNA polymerase, a small (~39 kDa) enzyme specialized for single-nucleotide gap–filling in base excision repair. Structurally, Pol β has an N-terminal 8 kDa domain with deoxyribose‑5′‑phosphate (dRP) lyase activity, and a C-terminal 31 kDa polymerase domain performs the nucleotidyl transferase reaction.[17] Unlike replicative polymerases, Pol β lacks a separate proofreading exonuclease domain, but instead relies entirely on its intrinsic selectivity at the insertion. In the unbound dNTP state, Pol β adopts an extended conformation, then the binding of a correct dNTP and DNA gap induces a large subdomain closure that precisely aligns the primer O3′, the incoming dNTP's α-phosphate, and two Mg^{2+} ions for catalysis.[33]

High fidelity in Pol β arises from an induced-fit mechanism and key active‑site contacts. Upon correct nucleotide binding, Pol β's N-subdomain swings in to form a closed complex over the nascent base pair. Then, the three active-site aspartates (Asp190, Asp192, Asp256) coordinate two Mg^{2+} ions and the primer terminus (O3′), enforcing correct geometry. If a mismatch is present, the closed conformation is destabilized: suboptimal base pairing leads to incomplete alignment and rapid reopening. Thus, as with B-family pols, Pol β achieves most discrimination at the binding/closure step. Its measured misincorporation frequency (~10^{-4}–10^{-6}) reflects both this selection and slower extension of mismatches.

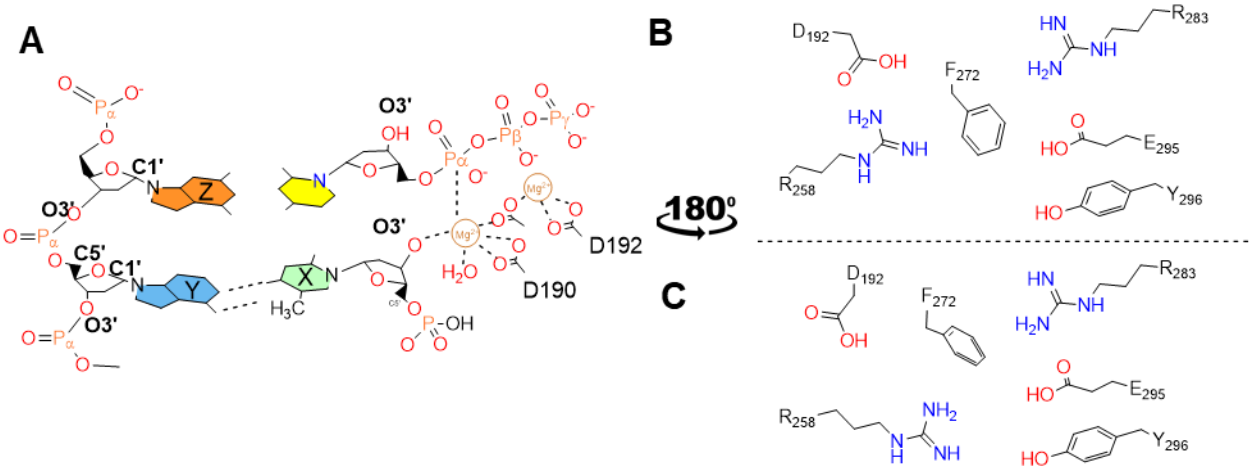


Figure 4. A) Schematic diagram of the nascent base pair (orange and cyan) showing the incoming nucleotide (yellow arrow) positioned for catalysis with the primer O3' aligned for nucleophilic attack. Key carbon positions (C1', C5') and O3' groups are labeled. Catalytic aspartates D190 and D192 coordinate the metal ions (Mg, tan spheres). (B–C) Two-dimensional interaction diagrams (rotated 180°) showing the hydrogen bonding network surrounding the catalytic site. Key residues D192, F272, R258, R283, E295, and Y296 form a network of interactions in the open and inactive (B) and the active state (C). The orientation of R258 and its interaction with D192 differs between conformational states, allowing the D192 coordinates the Metal center.

1.7. Mechanisms of correct nucleotide selection

As Pol's catalysis involves many steps, different mechanisms for correct nucleotide selection affect Pol's fidelity, depending on the structural architecture and its biological role. For example,

the initial binding of a deoxynucleoside triphosphate (dNTP) to the polymerase-DNA binary complex represents the first crucial step in nucleotide selection. This process is fundamentally governed by the geometric complementarity between the incoming nucleotide and the templating base, with a strong preference for Watson-Crick (WC) base pairing. The formation of correct hydrogen bonds and appropriate base stacking interactions associated with WC geometry energetically favors the binding of the dNTP. Additionally, upon dNTP binding, the transition from an "open" to a "closed" conformation creates a more tightly packed active site that sterically occludes mismatched nucleotides, which do not conform to the canonical WC geometry. The efficiency of this conformational closure is itself a selective filter, and the energy derived from correct WC interactions helps drive this transition more effectively than the weaker or distorted interactions formed by incorrect pairs, initiating the discrimination process even before the chemical step.[1,17]

The successful binding of a geometrically correct nucleotide, coupled with the subsequent induced-fit conformational changes, culminates in the precise preorganization of the polymerase active site for catalysis.[34] This critical preorganization involves the optimal alignment of the incoming dNTP, the primer terminus, essential catalytic metal ions, and key amino acid side chains that participate directly in the chemical reaction.[10] This meticulously arranged catalytic geometry is essential for efficient phosphodiester bond formation. Conversely, the binding of an incorrect or mismatched nucleotide typically fails to induce the same optimal conformational state. Instead, it often results in a misaligned or distorted active site geometry that is non-productive and energetically unfavorable for catalysis.[10] Such distortions can manifest as steric clashes, suboptimal hydrogen bonding patterns with active site residues, or improper coordination of the catalytic metal ions, all of which hinder the chemical reaction. Thus, the

initial geometric sensing of the incoming nucleotide is amplified through conformational adjustments, ultimately determining the catalytic competence of the active site.[35]

Ultimately, the chemical basis of polymerase fidelity resides in the enzyme's ability to differentially stabilize the transition state (TS) of the phosphoryl transfer reaction for a correct nucleotide compared to an incorrect one. This results in a lower activation energy for correct incorporation and a significantly higher activation energy for misincorporation. Central to this phenomenon is the concept of electrostatic preorganization, where the enzyme's active site provides an electrostatic environment that is preorganized to be complementary to the charge distribution of the transition state for the correct reaction.[6,36,37] This preorganized environment, established by the specific arrangement of active site residues, protein dipoles, and catalytic metal ions, minimizes the electrostatic penalty associated with charge rearrangement during the formation of the correct TS. When an incorrect nucleotide is present, the subtle geometric distortions it introduces, even if it passes earlier checkpoints, lead to a suboptimal fit within the active site. This disrupts the delicate electrostatic preorganization; the protein environment, including the orientation of its dipoles, and forces into a less effective configuration for stabilizing the incorrect TS, increasing the activation energy for misincorporation.

1.8. Ribonucleotide discrimination

The accurate synthesis of DNA requires polymerases to exhibit profound selectivity for deoxyribonucleoside triphosphates (dNTPs) over ribonucleoside triphosphates (rNTPs), despite the significantly higher cellular concentrations of rNTPs. This discrimination is critical for maintaining genomic integrity, as the misincorporation of ribonucleotides (rNMPs) renders the DNA phosphodiester backbone more susceptible to alkaline hydrolysis due to the presence of the

2'-hydroxyl group, potentially leading to strand breaks and complex cellular repair responses.[38] A predominant strategy employed by many DNA polymerases is the "steric gate" mechanism, wherein one or more amino acid residues within the enzyme's active site create a physical impediment that clashes with the 2'-hydroxyl of an incoming rNTP, thereby favoring dNTP binding. This steric exclusion can be mediated by an amino acid side chain or, in certain polymerases, by elements of the protein backbone.

The specific molecular strategies for ribonucleotide discrimination are well exemplified by DNA Polymerase β (Pol β) and the bacteriophage RB69 DNA polymerase (gp43), showcasing how different polymerase families have evolved distinct solutions. Pol β, a human X-family polymerase, excludes rNTP, causing unfavorable steric interaction between the 2'-hydroxyl of the incoming rNTP and the backbone carbonyl oxygen of Tyr271, located in the α-helix. Furthermore, the productive binding of a dNTP allows the hydroxyl group of Tyr271 to form a critical hydrogen bond with the minor groove edge of the primer terminus; this stabilizing interaction is disrupted or lost upon rNTP binding, contributing significantly to discrimination.[39]

RB69 polymerase, a B-family replicative enzyme, employs a more classical steric gate, the amino acid Tyr416. Computational studies, particularly free energy perturbation calculations, have indicated that the side chain of Tyr416 sterically occludes the 2'-hydroxyl group of an rNTP, thereby destabilizing its binding and favoring dNTP incorporation. Experimental validation through mutagenesis, such as replacing Tyr416 with a smaller alanine residue, significantly diminishes sugar selectivity, confirming the residue's pivotal role.[40]

1.9. Computational investigations

Computational research, employing methods like molecular dynamics (MD) and free energy calculations, has provided crucial atomistic insights into how DNA polymerases achieve fidelity. As shown before, the complex process of nucleotide addition involves the study of protein conformational changes, nucleotide binding, and reaction catalysis. Here, we emphasize relevant computational studies related to the functional role of the polymerase active site in regulating correct nucleotide selection and how computational methods can quantify variations in the binding energies of dNTP and rNTPs.

1.9.1.O3’ primer position and relevance in polymerase fidelity

The precise positioning of this O3' group, particularly its coordination to the catalytic metal ion MgA and alignment for in-line attack on the incoming dNTP's Pα atom, distinguishes catalytically 'reactive' from 'nonreactive' states. Previous results have shown that the ability of the polymerase active site to stabilize or destabilize the reactive O3' configuration, depending on the correctness of the incoming nucleotide, is a fundamental determinant of fidelity, varying significantly between enzymes with different accuracy levels.

For high-fidelity DNA polymerase β (Pol β), extensive MD and umbrella sampling/REST simulations showed the O3' alignment as an essential part of the Polymerase fidelity. These studies demonstrated that the reactive state, characterized by direct MgA-O3' coordination (~2.2 Å) and optimal O3'-Pα proximity (~3.3 Å), is energetically stable and preferentially populated only when the correct dCTP nucleotide is bound opposite template dG (favored by 3.3 kcal/mol). When the incorrect dATP is present, the simulations consistently show a lack of MgA-O3' coordination (replaced by water), leading to a nonreactive geometry with a much larger O3'-Pα distance (~5.9 Å), which is energetically favored (3.0 kcal/mol) over the strained reactive

alignment. This clear energetic distinction, directly linked to the critical MgA-O3' interaction, ensures that Pol β primarily adopts a catalytically competent geometry for correct nucleotides, thereby enforcing high fidelity.

Conversely, studies the low-fidelity DNA polymerase η (Pol η) uncovered a more permissive active site dynamic regarding O3' positioning. Free energy simulations indicated much smaller energy differences between reactive and nonreactive states for both matched (1.1 kcal/mol difference) and mismatched (2.0 kcal/mol difference) nucleotide pairs compared to Pol β. MD trajectories showed that even with the incorrect dATP nucleotide, the Pol η active site's flexibility allows it to dynamically access a reactive configuration (MgA-O3' ~2.2 Å, O3'-Pα ~3.2 Å) without significant structural penalties, closely mimicking the geometry of the matched complex. This ability to accommodate and catalytically align the O3' group for both correct and incorrect base pairs with relatively similar ease, stemming from less stringent structural constraints and shape complementarity requirements, fundamentally explains the reduced fidelity observed for Pol η. Thus, computation highlights that local, metal-mediated O3' dynamics is a critical, tunable factor controlling polymerase accuracy.[10]

1.9.2.Alchemical free energy for the study of polymerase catalysis

While prechemistry steps and the correct position of the catalytic center are crucial for polymerase's high accuracy, enzyme catalysis also plays an important role in Pols function. As the ratio of catalytic efficiencies defines fidelity, the study of Pol's accuracy involves the contribution of the chemical kinetics and the dNTP binding. Florian et al.[41] employed a Linear response approximation method to calculate the binding free energy of transition states of Pol beta, eta, y T7 DNA Pol, showing good agreement with steady-state kinetics data. These calculations for Pol β and analogous investigations for Pol η and T7 DNA Pol underscore that

the differential binding affinity of the correct versus incorrect dNTP to the enzyme-DNA complex, particularly at the transition state for the chemical incorporation step, is a crucial determinant of polymerase fidelity. More recent computational investigations have focused on the impact of distant single-point mutations on the catalytic efficiency of human Pol β. For instance, Klvana et al.[35] employed Linear Interaction Energy (LIE) techniques coupled with Free Energy Perturbation (FEP) through a thermodynamic cycle transforming wild-type into mutant enzymes. A key component of this transoformation is the presence of a shared dianionic phosphorane TS structure in both structures, allowing the calculation focus on the amino acid side change. This study found that introducing scaling factors (e.g., 0.5) to the FEP free energies, or using a hybrid FEP/LIE method where the LIE equivalent replaced the FEP van der Waals term for the mutated residue's interaction, significantly improved the agreement with experimental changes in fidelity upon mutation for both right and wrong dNTP insertions. Specifically, the scaled FEP calculations yielded results with standard deviations of 0.9 (for right dNTP, dC) and 1.1 (for wrong dNTP, dA) from the experimental log Ω values (relative insertion efficiencies). The hybrid FEP/LIE method further improved this, achieving standard deviations of 0.4 for the right dNTP and 1.0 for the wrong dNTP, demonstrating a closer match to experimental observations for several Pol β mutants like I174S, I260Q, and M282L. However, a persistent challenge in these MM-based approaches is the modeling of transition states, which often rely on non-standardized parameters that can vary significantly based on active site architecture and the number of catalytic metal ions considered. Furthermore, the assumption of an identical TS for wild-type and mutants, alongside the use of empirical scaling, represents inherent simplifications, and the convergence of FEP calculations remains computationally demanding.

QM/MM approaches, including methods like the Empirical Valence Bond (EVB) method, often combined with FEP and Umbrella Sampling (US), have been employed to address the chemical step with greater quantum mechanical detail. Early QM/MM work on T7 DNA polymerase, and subsequently on DNA Pol β, proposed a protein-mediated mechanism where a conserved aspartate residue (Asp654 in T7 Pol, Asp256 in Pol β) acts as the general base, abstracting the proton from the nucleophilic 3'-OH of the primer. This contrasted with alternative pathways involving proton transfer to bulk water or the α-phosphate of the incoming dNTP. Later QM/MM studies explored other possibilities; for example, Zhang and co-workers[33] proposed a water-mediated substrate-assisted (WMSA) mechanism for DNA polymerase IV (Dpo4), where a nearby water molecule accepts the proton and shuttles it to the pyrophosphate leaving group. Wang et al. further examined a similar water-mediated mechanism in Dpo4, involving two water molecules in the proton relay. Complementing these, Matute et al.[6] using EVB calculations for Pol β, also investigated the dichotomy of proton transfer, suggesting that PT to the bulk solvent could be a dominant mechanism for the wild-type enzyme and certain D256 mutants, potentially coexisting with the direct PT to D256. More recently, Genna, et al[9] proposed a "self-activated mechanism" (SAM) for nucleic acid polymerization, based on Car-Parrinello QM/MM simulations. A key feature of SAM is an intramolecular proton transfer from the 3'-OH of the incoming nucleotide directly to the pro-S oxygen of its β-phosphate, facilitated by a conserved intramolecular H-bond, leading to in situ deprotonation as nucleotide addition occurs. These diverse QM/MM studies underscore the complexity of the chemical step. While they offer a more refined description of bond-breaking and bond-making events, definitively elucidating the precise PT pathway and its exact energetics across different polymerases remains challenging,

the results is highly sensitive to the chosen computational protocol, the extent of conformational sampling.

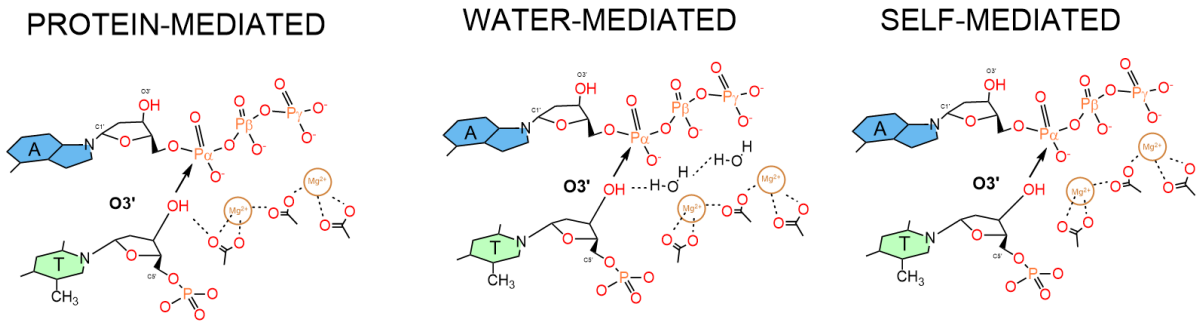


Figure 5 Catalytic modes of DNA polymerase phosphodiester bond formation. The nucleophilic attack of the primer 3'-OH on the dNTP α-phosphate is driven by three synergistic mechanisms: a, Protein-mediated: Conserved Asp residues coordinate two divalent metal ions. Metal A facilitates 3'-OH deprotonation, while Metal B stabilizes the pyrophosphate leaving group. b, Water-mediated: Structured water molecules act as proton wires or general bases to optimize 3'-OH nucleophilicity. c, Self-mediated: The dNTP substrate assists its own incorporation through triphosphate-mediated active site organization and hydrogen-bonding networks that stabilize the transition state

1.9.3.Alchemical free energy on ribonucleotide discrimination

Free energy studies have quantified the contributions of steric and electrostatic interactions to sugar selectivity. Warshel and Yoon[40] used empirical valence bond (EVB) and alchemical FEP simulations (with explicit "water flooding" treatment of internal waters) to analyze dCTP→rCTP discrimination in model polymerases. Their work confirmed that, in DNA polymerases, the 2′–OH–Tyr steric clash is the dominant factor: converting a dNTP into an rNTP raises the transition-state binding free energy mainly through increased van der Waals (steric) repulsion with the conserved tyrosine. By contrast, in RNA polymerases (where a ribose is natural), the corresponding tyrosine instead forms favorable hydrogen bonds with the 2′-OH. Thus, Warshel and Yoon found that steric effects explain most of DNA pol discrimination, whereas

electrostatics dominate in RNA pols. This conclusion is supported experimentally: for example, the Y416A mutant of RB69 DNA polymerase (analogous to Tyr271 in Pol β) incorporates rNTPs far more readily than wild type, demonstrating that loss of the steric gate tyrosine cripples sugar selectivity. Overall, both experiment and computation converge on a model where a conserved tyrosine enforces sugar fidelity by physically excluding a ribose 2′-OH, in line with Warshel's EVB-based free energy results.

1.9.4.Neural Network Potentials (NNPs) in Alchemical Free Energy Calculations: Improvements for Protein-Ligand Studies

While alchemical free energy calculations represent a powerful theoretical framework for studying biomolecular interactions, their accuracy is intrinsically linked to the fidelity of the underlying potential energy surface used to describe the system. These calculations have traditionally relied on classical molecular mechanics (MM) force fields for large protein-ligand complexes. MM force fields are computationally efficient, enabling the extensive sampling required for free energy convergence. However, they employ simplified functional forms and fixed atomic charges, which can limit their ability to accurately capture complex electronic phenomena critical for molecular recognition. These phenomena include electronic polarization, charge transfer between interacting molecules, which can significantly influence interaction energies and geometries, particularly in diverse chemical environments or near catalytic centers involving bond making or breaking. These inherent limitations of classical force fields can translate into inaccuracies in the calculated free energies, posing a persistent challenge in achieving high predictive power for protein-ligand binding affinities.[42]

The emergence of machine learning techniques, particularly Neural Network Potentials (NNPs), is imposing a paradigm shift in the development of more accurate and transferable potential energy surfaces for molecular simulations. NNPs are trained on extensive datasets derived from high-level quantum mechanical calculations, typically Density Functional Theory (DFT), and are designed to learn the complex relationship between atomic coordinates and the system's energy and forces. Once trained, NNPs can predict these properties with an accuracy that approaches that of the reference QM method but at a lower computational cost, making them compatible with molecular dynamics simulations. For applications to large biomolecular systems, hybrid NNP/MM methodologies have been developed. In these schemes, the NNP is used to describe a chemically important region of the system with high QM-like accuracy. At the same time, the remaining of the protein and solvent are treated with a computationally less demanding classical MM force field.[42] This approach effectively balances the need for accuracy in the region of interest with overall computational efficiency. Crucially, these advanced NNP and NNP/MM potentials are being progressively integrated into alchemical free energy frameworks, like the Alchemical Transfer Method (ATM), to elevate the accuracy and reliability of binding free energy predictions for protein-ligand systems. Several recent studies have demonstrated improvements in the accuracy of relative binding free energy calculations when using NNP/MM potentials compared to those obtained with conventional MM force fields like GAFF2.[43]

While much of the developmental effort and application focus has been on protein-ligand binding for drug discovery, the enhanced accuracy of NNPs holds significant potential for advancing the study of more complex enzymatic systems, including polymerases. By providing more faithful and nuanced representations of enzyme-substrate and enzyme-inhibitor

interactions, NNP-based alchemical calculations could offer a more reliable quantification method for the large variety of substrates that polymerases process.

1.10. Applications and relevance

1.10.1. Genome stability and mutagenesis

DNA polymerases are responsible for genomic integrity, playing an indispensable role in maintaining genome stability through the accurate replication and repair of DNA. High-fidelity DNA polymerases, in particular, are responsible for the precise duplication of the genetic blueprint during cell division, a crucial process for the faithful transmission of genetic information to subsequent generations.[2] Despite all the safeguards previously mentioned, errors in polymerase function, such as the misincorporation of nucleotides that evade proofreading, or deficiencies in polymerase-associated DNA repair pathways, can lead to mutations. While the frequency of such base substitution errors is typically low, ranging from 10^{-3} to less than 10^{-6} per incorporated nucleotide, the vast size of genomes means that even these low error rates can result in a significant mutational load over time.[1] The accumulation of these mutations and the resultant genomic instability lead to various human diseases, most notably cancer, where variants of certain DNA polymerases, such as DNA polymerase β, found in tumor cells often exhibit altered fidelity or catalytic activity and can contribute to cellular transformation and malignancy.[17] This underscores that genome stability is not a passive state but a dynamic equilibrium, actively maintained by a complex network of high-fidelity replication and multifaceted repair processes, where polymerases are central players.

The cellular genome is constantly exposed to endogenous metabolic byproducts and exogenous environmental agents that can cause a wide array of DNA lesions. Such damage, if unrepaired,

can pose significant barriers to the progression of high-fidelity replicative polymerases, potentially leading to stalled replication forks, chromosomal aberrations, and ultimately, genomic instability.[44] To counteract these threats, organisms have evolved specialized DNA polymerases, prominently including members of the Y-family (such as Pol η, Pol ι, Pol κ) and others like Pol ζ (a B-family member), which are capable of performing translesion synthesis (TLS), which is a "damage-bypass" mechanism that allows the replication machinery to skip over DNA lesions rather than stopping, ensuring cell survival at the risk of introducing mutations. These TLS polymerases typically possess more spacious and flexible active sites, enabling them to accommodate distorted DNA templates or bulky lesions that would otherwise block replicative enzymes.[45] While TLS is crucial for completing DNA replication and ensuring cell survival in the face of persistent DNA damage, it often comes at a cost: TLS polymerases generally exhibit lower fidelity than their replicative counterparts, not only when copying damaged DNA but also on undamaged templates. This inherent propensity for error makes TLS a significant source of mutagenesis. Consequently, the recruitment and activity of TLS polymerases must be meticulously regulated, often involving intricate polymerase switching mechanisms, to ensure they are employed only when essential and for minimal stretches of synthesis, showing a delicate balance between genome duplication and the preservation of genetic integrity.[46]

Additionally, the Base Excision Repair (BER) pathway is another DNA repair mechanism focused on a wide array of small, non-helix-distorting base lesions, such as those induced by oxidation, deamination, and alkylation.[47] DNA polymerase β (Pol β), a key member of the X family of DNA polymerases, is central to BER. The BER process begins when a DNA glycosylase recognizes and removes the damaged base, creating an apurinic/apyrimidinic (AP) site. An AP endonuclease subsequently cleaves this site, preparing it for Pol β.[48] Then, its 8 kDa

N-terminal domain, which has 5'-deoxyribose phosphate (dRP) lyase activity, removes the dRP moiety, and its 31 kDa C-terminal domain exhibits polymerase activity to fill the single-nucleotide gap accurately.[49] Finally, DNA ligase seals the nick, restoring DNA integrity. While the core enzymatic steps of Pol β in BER are well-defined, key questions remain about the precise in vivo coordination of the BER machinery and the mechanisms regulating Pol β activity and recruitment under varying lesion contexts or cellular stresses.[50] Moreover, understanding the impact of numerous Pol β cancer-associated variants and delineating its non-canonical roles beyond short-patch BER are critical research areas to fully map its contributions to genome maintenance and disease.

1.10.2. Polymerase inhibitors in antiviral and anticancer drug design.

The targeted inhibition of DNA polymerase activity offers a potent strategy against diseases characterized by uncontrolled or unwanted cellular proliferation, such as cancer, and against infections caused by viral and bacterial pathogens.[4] The rationale for this approach lies in the essentiality of polymerases for DNA replication and repair, processes upon which rapidly dividing cancer cells and replicating pathogens heavily rely. A prominent class of polymerase inhibitors comprises nucleoside analogs (NAs), which are structurally modified versions of natural nucleosides. Following administration, NAs undergo intracellular phosphorylation to their active triphosphate forms, acting as fraudulent substrates for DNA polymerases.[51] Their incorporation into the nascent DNA strand typically leads to chain termination, as they often lack the crucial 3'-hydroxyl group required for further phosphodiester bond formation or possess sugar modifications that sterically hinder extension. Clinically important examples include treatment regimens against retroviruses like HIV (e.g., zidovudine[52] targeting reverse

transcriptase), herpesviruses (e.g., acyclovir targeting viral DNA polymerase)[53], and other significant viral pathogens.[11,51] Similarly, in oncology, NAs such as gemcitabine and fludarabine are extensively used in chemotherapy to exploit the heightened replicative demands of cancer cells.[54] Another major group is the non-nucleoside inhibitors (NNIs), which, unlike NAs, do not mimic natural nucleosides and bind to allosteric sites on the polymerase, remote from the catalytic center.[55] This binding event typically induces conformational changes in the enzyme that impair its catalytic function, often by disrupting the precise alignment of substrates or by hindering essential domain movements, without directly competing with natural dNTP binding. Nevirapine, an anti-HIV drug, exemplifies this class.[56]

Emerging therapeutic strategies are also exploring the inhibition of specific DNA repair or translesion synthesis (TLS) polymerases. For example, DNA polymerase θ (Pol θ), which is often overexpressed in certain cancers and plays a role in alternative DNA repair pathways, is a promising target, particularly for overcoming resistance to conventional therapies.[57] Similarly, DNA polymerase η (Pol η), another key translesion synthesis enzyme, plays a significant role in chemoresistance by enabling cancer cells to bypass DNA lesions induced by platinum-based drugs like cisplatin.[58] Consequently, the development of Pol η inhibitors is actively being pursued as a strategy to sensitize cancer cells to existing chemotherapies and overcome drug resistance, with recent computer-aided drug design efforts leading to the identification of novel classes of such inhibitors.[59] Beyond these established areas, bacterial DNA polymerase I and parasitic polymerases, like the apicoplast-localized enzyme in Plasmodium falciparum, are being investigated as targets for novel antibiotics and antimalarial drugs, respectively.[4,55]

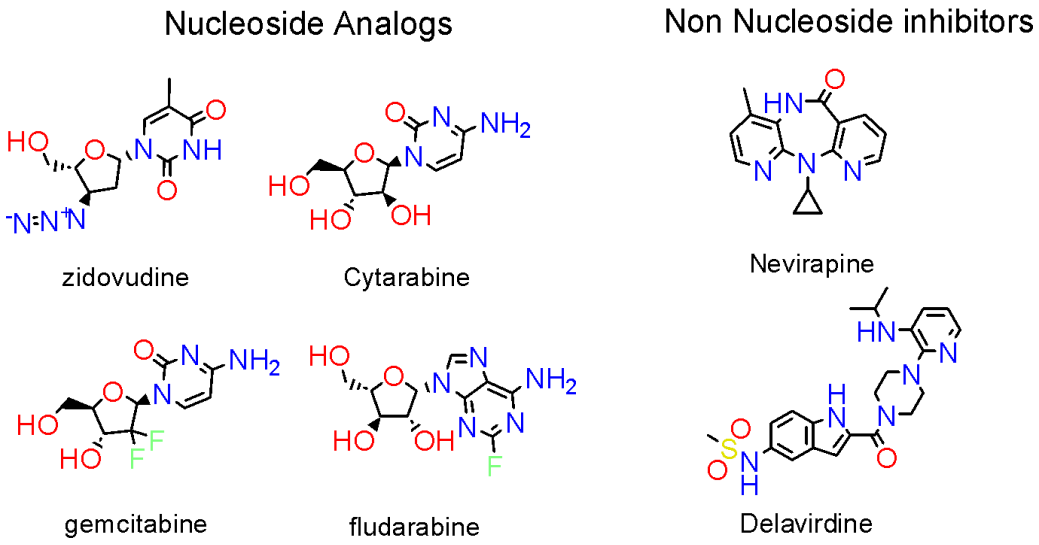


Figure 6. (Left) Nucleoside analog inhibitors including zidovudine (anti-HIV), cytarabine (anticancer), gemcitabine (anticancer), and fludarabine (anticancer) that act as chain terminators or competitive substrates. (Right) Non-nucleoside inhibitors nevirapine and delavirdine (anti-HIV) that bind to allosteric sites remote from the catalytic center, inducing conformational changes that impair polymerase function without competing with natural nucleotide substrates.

1.10.3. Biotechnological applications of modified nucleotides

DNA polymerases routinely interact with a diverse spectrum of modified nucleotides, a category that encompasses not only naturally occurring DNA lesions—such as oxidized bases like 7,8-dihydro-8-oxoguanine (8-oxoG), abasic sites, or formamidopyrimidine (Fapy) lesions—but also a vast array of synthetic nucleotide analogs designed for research or therapeutic purposes.[12] The study of these interactions is fundamental to understanding the intricacies of polymerase function. For instance, investigating how polymerases process DNA lesions provides critical insights into DNA repair pathways and the origins of mutagenesis. Similarly, synthetic analogs, such as non-hydrolyzable dNTPs that trap catalytic intermediates or fluorescently labeled nucleotides like 2-aminopurine that report on conformational changes, serve as invaluable probes for dissecting polymerase catalytic mechanisms, substrate specificity, fidelity checkpoints, and dynamic structural transitions.[60]

The interplay between DNA polymerases and modified nucleotides has been ingeniously exploited to develop a multitude of transformative biotechnological applications that are now central to molecular biology research and diagnostics.[1,12,60] Perhaps the most prominent example is DNA sequencing, where next-generation sequencing (NGS) technologies overwhelmingly rely on the ability of DNA polymerases to incorporate specifically modified nucleotides, such as those carrying fluorescent labels or reversible terminator groups, enabling massively parallel sequence determination.[61] The polymerase chain reaction (PCR) and various molecular diagnostic assays also extensively utilize modified nucleotides for purposes such as DNA labeling, the introduction of specific mutations, or real-time detection of amplification products; polymerases engineered for enhanced tolerance to these modifications or to PCR inhibitors found in clinical samples are crucial for robust and sensitive diagnostics.[62] Furthermore, synthetic biology is increasingly dependent on engineered polymerases that can synthesize and replicate xeno-nucleic acids (XNAs)—artificial genetic polymers possessing non-natural sugar backbones or nucleobases. These XNAs can exhibit novel properties, such as enhanced chemical stability, resistance to nucleases, or unique catalytic activities, opening avenues for new therapeutics (e.g., XNA aptamers), advanced diagnostics, and even high-density information storage.[63] The development of these diverse biotechnologies frequently necessitates the sophisticated engineering of DNA polymerases, through techniques like directed evolution or structure-guided rational design, to optimize their efficiency, fidelity, and specificity for incorporating a wide range of modified or entirely unnatural nucleotide substrates.[60,64]

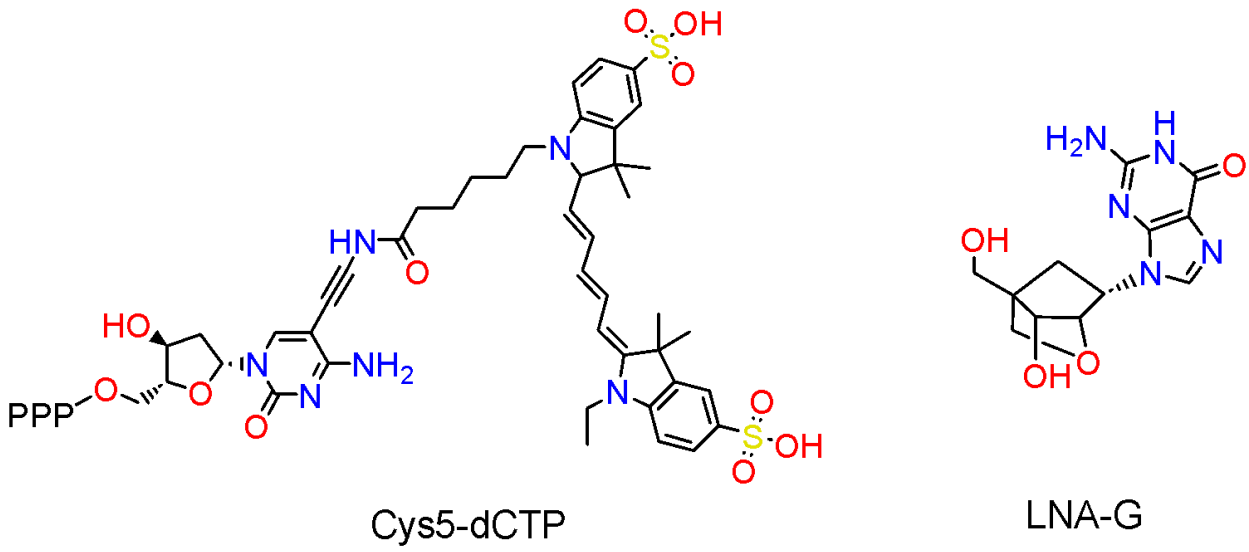


Figure 7 Chemical structures of representative modified nucleotides: Cys5-dCTP, a fluorescently labeled cytosine nucleotide conjugated with Cy5 dye via a flexible linker, used for fluorescence-based DNA sequencing. LNA-G (locked nucleic acid guanosine), featuring a methylene bridge connecting the 2'-oxygen and 4'-carbon of the ribose ring, which constrains the sugar in a C3'-endo conformation and enhances binding affinity to G-quadruplexes.

1.11. Computational methods for molecule generation in Structure Based Drug Design (SBDD)

Structure-based Drug Design aims at an accurate understanding of the target's binding site architecture, and its potential interaction points can guide the generation of complementary molecular structures. Molecular modelling techniques such as molecular dynamics simulations and free energy calculations provide useful insights into protein dynamics and the quantification of binding affinity in a wide range of protein-complex systems.[65] However, the design of novel molecules that fulfill different pharmacokinetic criteria, strong affinity, and high selectivity can not be solved efficiently using only these techniques. In addition to these methods, SBDD involve key computational tasks such as the identification and characterization of binding sites, the design of novel molecular entities (de novo) or relevant chemical fragments to link into an

existing scaffold (Fragment Based Drug Design), and a fitness score that allows to iterate the process continuously.[66]

The computational tools for molecule construction, such as growing, linking, and merging, exhibit significant overlap between *de novo* Structure based drug design and fragment based drug design. However, de novo design typically begins with an empty binding site or a very rudimentary seed, whereas FBDD starts with experimentally validated or high-confidence computationally identified fragment hits. This difference means that FBDD often operates within a more constrained search space for these elaboration strategies, guided by the known interactions of the initial fragments. Nevertheless, the underlying algorithmic principles for extending or combining molecular moieties share a common foundation.[67]

The advancement and broader applicability of both *de novo* SBDD and FBDD are based on the increasing availability of high-quality 3D structural data, both from experimental sources and from breakthroughs in protein structure prediction like AlphaFold.[68] This wealth of structural information fuels the algorithms used in SBDD, allowing for more targets to be addressed and potentially improving the reliability of structure-based predictions as models can be trained on larger and more diverse datasets.[69]

Regardless of whether the strategy is purely de novo or involves the elaboration of identified fragments, ensuring the synthetic accessibility of the computationally designed molecules remains a critical and persistent bottleneck. Bridging the gap between theoretically optimal molecular structures and those that can be practically synthesized in the laboratory is an ongoing challenge that computational methods are continuously striving to address more comprehensively.

1.11.1. Deep learning in the context of SBDD

Deep learning (DL), a specialized subfield of machine learning, has emerged as a transformative technology across numerous scientific disciplines, including pharmaceutical research. Deep learning algorithms are characterized by the use of artificial neural networks (ANNs) with multiple layers (hence "deep" architectures). These layers enable the hierarchical extraction and learning of complex patterns and abstract representations directly from raw input data. The capacity of these models to automatically learn relevant features, often obviating the need for extensive manual feature engineering, is one of their significant advantages. In the context of drug discovery and development, its potential is being increasingly recognized and exploited across a wide spectrum of applications, including target identification and validation, molecular property prediction (e.g., ADMET, bioactivity), de novo molecular design, and chemical synthesis planning.[70,71]

The integration of deep learning into SBDD workflows is providing novel and powerful solutions to some of its most pressing challenges. Deep learning models are being developed to enhance the accuracy of binding affinity prediction, improve the efficiency and effectiveness of virtual screening campaigns, and facilitate the generation of novel molecular structures specifically tailored to bind target protein pockets with high affinity and selectivity. Geometric deep learning, focuses on processing data with underlying geometric structures (such as 3D molecular coordinates or graphs), is particularly pertinent to SBDD, as it can directly learn from the spatial and relational information inherent in protein-ligand complexes. By training on large datasets of known protein-ligand interactions and chemical structures, these models can capture patterns and relationships that may not be apparent through traditional physics-based or empirical approaches, potentially leading to more accurate predictive models and more innovative generative strategies.[72]

A critical aspect of applying deep learning to SBDD is the effective representation of 3D protein environments. Voxelization, which discretizes 3D space into a grid of volumetric pixels (voxels), has emerged as a common and effective technique, particularly for Convolutional Neural Networks (CNNs) that are adept at processing grid-like data. For instance, GNINA[73] or KDeep[74] leverages CNNs, often with voxelized inputs representing the protein binding site and ligand, for tasks such as molecular docking, pose prediction, and binding affinity scoring. These methods underscore the power of encoding detailed 3D structural information to inform ML-driven drug design.

The remarkable capabilities of deep learning are intrinsically linked to the availability, quality, and relevance of the data used for training these models. While complex models excels at learning from large datasets, its performance can be significantly affected by data scarcity, biases present in the training data, or when applied to chemical or biological domains that are poorly represented in existing databases or with experimental measurements with great variability and uncertainty. This data dependency is a critical consideration in the practical application of deep learning approaches into SBDD challenges.

1.11.2. Molecule generation as a classification problem

Deep learning models for molecular generation within SBDD can serve two distinct purposes: regression models that predict continuous binding affinity scores for ranking pre-existing molecular libraries, or end-to-end generative models (such as Variational Autoencoders or Generative Adversarial Networks) that learn probability distributions over molecular space to directly sample novel molecules.[75] Parallely, certain approaches involve training a neural network to solve a classification problem, where the output of this classification task directly

informs or guides the construction, selection, or placement of a molecule or its constituent fragments within a protein's binding site. In these strategies, the classification step is an integral part of the molecule generation or assembly process itself.

The "classification" can take various forms: it might involve predicting a specific class or type of fragment suitable for a given context, determining the presence or absence of particular chemical features or substructures, or assessing the "correctness" or "appropriateness" of a proposed molecular component based on the surrounding protein environment and any existing ligand scaffold. The outcome of this classification then drives subsequent decisions in the molecular design workflow. DeepFrag[76] stands out as a prime example of this classification-driven strategy. It employs a 3D CNN that processes a voxelized grid representation of the protein receptor and the initial part of the ligand (the "parent" molecule). The model is trained to classify which fragment, from a library of thousands, is the correct one to reconstruct an original, known ligand from which the fragment was notionally removed. It has shown a strong ability to identify the correct fragment. It often suggests chemically similar and plausible alternatives when the exact fragment is not its top prediction, making it a valuable tool for lead optimization. Specifically, DeepFrag predicts a molecular fingerprint (specifically, an RDKFingerprint vector) that describes the desired missing or new fragment. This predicted fingerprint is then used to query a large library of over 6500 known, pre-existing fragments, which are ranked based on their cosine similarity to the predicted fingerprint. This classification-like or ranking approach has several inherent strengths. Firstly, because the suggested fragments are drawn from a library of known chemical entities, their chemical validity is assured, a significant advantage over some generative models that may produce syntactically or chemically incorrect structures. Secondly, this formulation allows for a more straightforward and quantitative evaluation of the model's

performance using metrics such as TOP-k accuracy (i.e., the frequency with which the correct fragment is found among the top k predictions). Thirdly, while the raw fingerprint vector is not directly human-interpretable, the final output, a ranked list of known chemical fragments, is readily understandable by medicinal chemists.

Other methodologies also adopt or incorporate classification-like principles for fragment-based design. The FRAME (Fragment-Based Molecular Expansion) framework,[77] involves a broader system for iterative ligand expansion, which includes an “Attachment location model” for the selection of suitable positions for fragment growing and, also a "Fragment Scoring Model" that selects the most suitable fragment to add from a library and predicts its geometry; this selection from a discrete set based on learned scores can also be considered a classification or ranking task.

These classification-driven approaches benefit from leveraging libraries of known, often synthetically tractable fragments and directly utilize the 3D context of the protein pocket. However, their performance is intrinsically linked to the scope and quality of the fragment libraries and training datasets, and they may be more tailored to lead optimization scenarios rather than discovering entirely novel chemical scaffolds from first principles.

2. Theory and Principles

2.1. Molecular dynamics

Molecular Dynamics (MD) simulations allow the observation of the temporal evolution of molecular systems with atomic-level resolution. By numerically integrating Newton's classical equations of motion, MD simulations generate trajectories that describe the positions and velocities of atoms over time. From these trajectories, macroscopic thermodynamic and kinetic properties of the system can be derived, contributing to the deeper understanding of molecular structure, dynamics, and function, and therefore proving particularly valuable for the study of complex biomolecular systems. The basic premise involves calculating the forces acting on each atom based on a defined potential energy function and then using these forces to propagate the atomic positions and velocities at a subsequent small time increment.[65,78]

$$F_i(t) = m_i a_i = -dV(x(t))/dx_i(t) \# (1)$$

The integration algorithm and the length of the time step represent a critical balance between computational efficiency and the numerical stability and accuracy of the simulation. Several numerical algorithms have been developed for this, with the Verlet algorithm and its variants, such as the leap-frog algorithm, being among the most commonly employed.[79] For instance, the

basic Verlet algorithm calculates the position of an atom at time t+Δt using its positions at time t and t−Δt, and its acceleration at time t, according to the formula:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + a(t)(\Delta t)^2 \#(2)$$

Where the acceleration $a(t)$ can be found using equation (1). At the beginning of the MD, the positions $r(t - \Delta t)$ can be estimated by the Taylor series expansion:

$$r(-\Delta t) \approx r(0) - v(0)\Delta t + \frac{a(0)}{2}(\Delta t)^2 \#(3)$$

Where the velocity at the first time step matches the average kinetic energy that corresponds to a desired simulation temperature. In contrast, algorithms such as leapfrog update positions and velocities at alternating half-time steps, which offer good numerical stability and efficiency, while calculating explicitly the velocity at a given timestep. These types of algorithms are favored due to their time-reversibility and excellent long-term energy conservation properties. The time step (Δt) must be chosen small enough to resolve the highest frequency motions present in the system, typically around 1 fs for hydrogen bond vibrations. If Δt is excessively large, these rapid motions are not integrated accurately, potentially leading to numerical instability and unphysical behavior. To mitigate this and allow for slightly larger time steps (e.g., 2 fs), constraint algorithms such as SHAKE or LINCS[80] are often utilized. These algorithms fix the lengths of bonds involving light atoms like hydrogen, effectively removing the fastest vibrational modes from the system and thereby permitting a larger integration time step without compromising the stability of the simulation for the slower, often more functionally relevant, motions. This represents a practical compromise to enhance computational efficiency.[65,81]

To better mimic actual macroscopic behavior, specific algorithms such as thermostats and barostats are used to regulate temperature and pressure. Generally, these algorithms try to

regulate a macroscopic property by alterations in the integration of motion that lead to desired ensembles, such as NPT, where the number of molecules, pressure, and temperature are constant. Specifically, Thermostats alter the equation of motion in a deterministic or stochastic manner, rescaling the velocity after the particles' positions and momenta are updated by the integrator. Similarly, barostats scale the system volume during the integration of Newton's equations, which allows the regulation of the pressure.[81]

2.2. Force Fields

The potential energy $V(x(t))$ of a molecular system in equation (1), from which forces are derived, is described by an empirical function known as a molecular mechanics force field. These functions are collections of equations and associated parameters that approximate the true quantum mechanical potential energy surface of the system using classical physics expressions. AMBER (Assisted Model Building with Energy Refinement) and CHARMM (Chemistry at HARvard Macromolecular Mechanics) FF are widely used functional forms, where the total potential energy is decomposed into contributions from bonded and non-bonded interactions.[82–84] A general representation of such a force field can be written as:

$$V = \sum^{bonds} \frac{K_b}{2}\left(b - b_0\right)^2 + \sum^{angles} \frac{K_\theta}{2}\left(\theta - \theta_0\right)^2 + \sum_{i}^{torsions} \left\{ \sum^{M} V_{ik}\left[1 + \cos cos\left(n_{ik}\phi_{ik} - \phi_0\right)\right] \right\} + \sum_{i,j}^{pairs} \epsilon_{ij}\left[\left(\frac{r_0}{r_i}\right.\right.$$

The bonded terms describe interactions between atoms that are covalently linked and include Bond, Angle, and Dihedral terms. Bond stretching terms model the energy required to stretch or compress a covalent bond from its equilibrium length b_0. This is typically represented by a harmonic potential, where K_b is the bond force constant. Angle bending terms account for the energy associated with deforming a bond angle θ from its equilibrium value θ_0, also commonly modeled harmonically, with K_θ being the angle force constant. Dihedral (or torsional) terms

describe the energy profile associated with rotation ϕ around a chemical bond. This is usually represented by M periodic Fourier series, where V_{ik} is the barrier height, n_{ik} is the multiplicity, and ϕ_0 is the phase angle for each component of the torsion. Some force fields, like CHARMM, also include Urey-Bradley terms (a cross-term linking bond stretching and angle bending for atoms separated by two bonds) and improper dihedral terms. Improper dihedrals are essential for maintaining the planarity of groups like aromatic rings or the chirality of stereocenters.[79,82]

The non-bonded terms describe interactions between atoms that are not directly bonded, typically those separated by three or more bonds. These terms include van der Waals and E=electrostatic interactions**.** The former accounts for short-range repulsion and long-range dispersion (attraction). These are most commonly modeled using the Lennard-Jones (LJ) 6-12 potential, where ε_{ij} is a parameter defining the depth of the energy well (strength of the interaction) and r_0 describes the minimum energy distances for the atom pair ij. The latter one describes the Coulombic forces between atoms. These are calculated using Coulomb's law, where q_i and q_j are fixed partial atomic charges assigned to atoms i and j, r_{ij} is the distance between them, ϵ_0 is the permittivity of free space, and ϵ is the dielectric constant (often set to 1 in explicit solvent simulations with non-polarizable models).[79,82]

The process of parameterization involves determining the values for all the constants in the force field, including reference distances, angles, charges, etc. This is a complex and iterative procedure where parameters are fitted to reproduce a wide range of experimental data (such as molecular geometries from X-ray crystallography or electron diffraction, vibrational frequencies from IR/Raman spectroscopy, and bulk thermodynamic properties like heats of vaporization and

liquid densities) and/or high-level quantum mechanical (QM) calculations performed on small, representative model compounds. Parameters are assumed to be applicable when the same chemical moieties (defined by atom types) appear in much larger and more complex molecules. However, this assumption can lead to inaccuracies if the local chemical or electrostatic environment of an atom type in a large, complex system significantly deviates from the environments present in the small model compounds used for its parameterization. Additionally, atomic partial charges in FF are fixed, meaning they do not respond to changes in the local electrostatic environment. This can be a significant approximation, particularly when simulating systems in heterogeneous dielectric environments (like a protein-water interface), interactions involving highly charged species, or processes where charge redistribution plays a critical role.[85] Due to this transferability problem, different strategies for parametrization are developed to address the chemical character of different types of molecules. For example, AMBER force fields are developed with specific parameters for proteins (e.g., ff19SB)[86,87], DNA (e.g., OL21)[88], RNA (e.g., OL3)[89], lipids (e.g., lipid21)[90], and carbohydrates (e.g., GLYCAM)[91].

For the accurate simulation of small organic molecules, often ligands or drug-like molecules, the General AMBER Force Field (GAFF) and its successor, GAFF2, were designed for broad compatibility with the AMBER family of force fields for proteins and nucleic acids. GAFF parameterization involves specific atom typing based on element, hybridization, and chemical environment, with bonded parameters derived from quantum mechanical (QM) calculations, empirical rules, and crystal structures; atomic charges are typically derived using the Restrained Electrostatic Potential (RESP) method at the HF/6-31G* level or the semi-empirical Austin Model 1 with Bond Charge Corrections (AM1-BCC).[85] While RESP charges offer higher

accuracy, their computational expense has led to the widespread practical use of AM1-BCC, a trade-off between rigor and throughput often encountered in force field development. GAFF2 represents an iterative improvement, featuring updated bonded parameters from higher-level QM on more compounds and refined non-bonded parameters for better reproduction of ab initio interaction energies and liquid properties. Other widely used small molecule force fields include the CHARMM General Force Field (CGenFF), which employs a rules-based penalty score for parameter assignment by analogy and a charge increment scheme, with its latest version (CGenFF v5.0) trained on an expanded set of 1390 molecules to improve intramolecular geometries and interaction energies.[92] The OPLS-AA (Optimized Potentials for Liquid Simulations - All-Atom)[93] force field derives its parameters, particularly torsional and nonbonded terms, by fitting to QM rotational energy profiles and experimental liquid properties, ensuring robust performance for organic liquids. The continuous evolution of these general force fields, driven by increasing computational power and larger validation datasets, underscores a trend towards greater accuracy and broader applicability, crucial for their role in complex biological simulations.

For modelling of DNA and RNA, the development of force fields has been particularly challenging due the presence of highly charged phosphate backbone, numerous flexible dihedral angles (α, β, γ, δ, ε, ζ), variable sugar puckering, complex glycosidic torsion (χ) behavior, and critical interactions with metal ions and the solvent environment (See Figure 7).[94] Key DNA force fields include parmbsc1, which refined ε/ζ and χ torsions for B-DNA, OL15, which focused on the β dihedral, and OL21, which improved α/γ torsions enhancing Z-DNA modeling. On the other hand, early improvements for RNA FF included the parmbsc0 correction for α/γ torsions and the influential χOL3 correction, which addressed the overestimation of syn base

conformations and improved the syn-anti balance. A particularly impactful development, originating from D.E. Shaw Research by Tan et al,[95] involved an extensive revision of AMBER ff14. This work achieved a significant step towards protein-level accuracy by re-parameterizing nucleobase nonbonded parameters (charges and van der Waals terms) against high-level QM base-pairing and stacking energies, alongside refining χ, γ, and ζ torsional potentials. A notable development is the DES-Amber DNA force field, which successfully transferred nonbonded parameters from the corresponding DES-Amber RNA force field, requiring only DNA-specific adjustments to χ and ζ torsions. This highlights the potential for efficient co-development leveraging chemical similarities between RNA and DNA. DES-Amber demonstrated good performance for ssDNA flexibility. Despite these advances, current force fields exhibit system-dependent accuracy, and modeling large conformational changes, ion interactions (especially Mg^{2+}) with full quantum effects, and the complex cellular milieu remain formidable challenges, suggesting that a universal RNA force field is yet to be realized. [94]

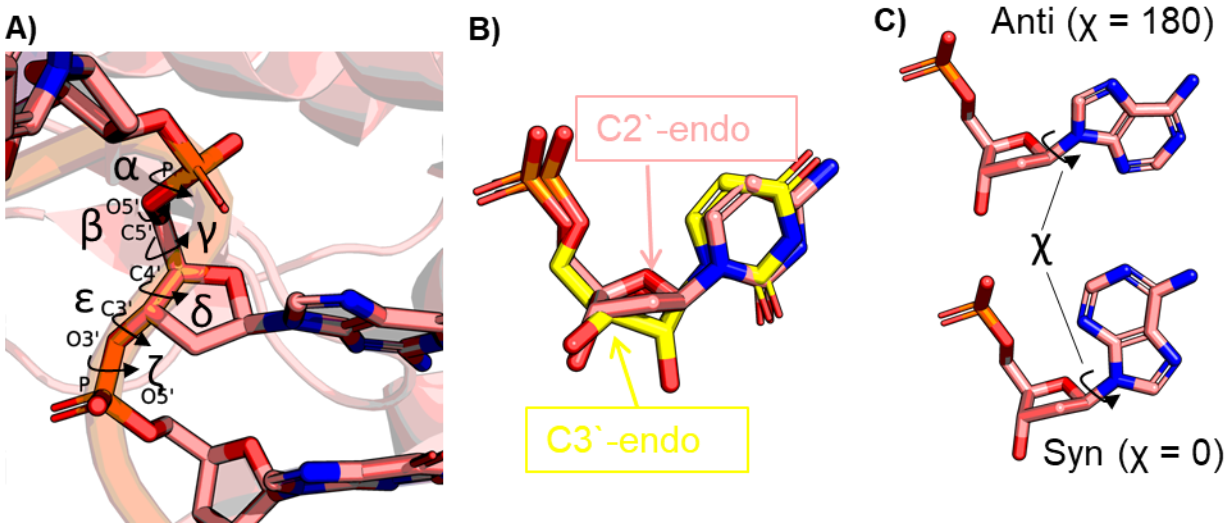


Figure 8. A) DNA-RNA phosphate backbone showing the six torsional angles (α, β, γ, δ, ε, ζ) that define backbone flexibility. The O3' and O5' atoms connect adjacent sugar moieties through phosphodiester bonds. B) Sugar pucker conformations comparing DNA (salmon, C2'-endo pucker) and RNA (yellow, C3'-endo pucker). The presence of the 2'-OH group in RNA favors

the C3'-endo conformation, while DNA adopts the C2'-endo pucker. C) Glycosidic torsion angle (χ) showing anti ($\chi = 180°$) and syn ($\chi = 0°$) conformations of the nucleobase relative to the sugar ring, which influence base stacking and overall helix geometry.

2.3. Neural Networks Potentials

Classical molecular mechanics force fields, while computationally efficient and widely used, have inherent limitations in terms of accuracy and transferability. These limitations arise primarily from their reliance on fixed, relatively simple functional forms and parameterization strategies that often struggle to capture complex quantum mechanical effects or adapt to diverse chemical environments. In recent years, Neural Network Potentials (NNPs) have emerged as an alternative. NNPs aim to achieve accuracies approaching that of QM calculations for potential energies and forces, but at a computational cost significantly lower than direct QM methods, although typically higher than that of classical force fields. The fundamental approach of NNPs is to learn the high-dimensional relationship between the spatial arrangement of atoms in a molecule or material and its potential energy directly from large datasets generated by QM calculations.[96]

For a NNP to be physically meaningful and generalizable, it must rigorously adhere to fundamental physical symmetries. The total energy of a system must exhibit permutation invariance, meaning it remains unchanged when identical atoms are exchanged. Similarly, the energy must be invariant to rigid translation and rotation of the entire system. From these energy invariances, it follows that atomic forces, as vector quantities, must be rotationally equivariant, meaning they must rotate in precisely the same manner as the system itself. The explicit incorporation of these symmetries within the NNP architecture is of paramount importance. It not only enhances data efficiency, allowing models to learn effectively from smaller datasets, but

also improves accuracy, robustness, and the ability to generalize to unseen molecular configurations.[97]

Many foundational and widely adopted NNP architectures are based on the scheme originally proposed by Behler and Parrinello.[98] In this approach, the total potential energy of a system is expressed as a sum of individual atomic energy contributions:

$$E_{total} = \sum_i E_i \#(5)$$

Where E_i is the energy contribution of atom i. Each atomic energy is predicted by a separate feed-forward neural network. Crucially, the parameters of this neural network are specific to the chemical element of atom i. The input to each atomic neural network is not the raw Cartesian coordinates of the atoms, but rather a feature vector, often called a descriptor or fingerprint, that quantitatively describes the local chemical environment of atom i up to a defined cutoff radius.

The ANAKIN-ME (ANI) family of potentials, developed by Smith, Isayev, and Roitberg[99], utilizes a refined version of Behler-Parrinello symmetry functions to construct what they term Atomic Environment Vectors (AEVs). These AEVs are specifically designed to be more transferable across diverse organic molecules by creating a feature space where common chemical motifs (like functional groups or ring structures) have recognizable signatures. The AEV for each atom is a vector composed of numerous radial and angular symmetry function values, calculated for all relevant pairs and triplets of atom types within the cutoff radius. Additionally, the result vectors ensure invariance to permutation, translation, and rotation. While this approach is computationally efficient and has shown high accuracy for specific domains like small organic molecules, its primary limitation lies in its use of "hand-crafted," fixed descriptors.

This rigidity can restrict the model's expressiveness and transferability to new chemical environments, and its reliance on a local cutoff can be a source of error in systems dominated by long-range interactions.

In contrast, other NNPs, such as TensorNet,[100] explicitly incorporate more complex representations of atomic environments. In detail, it processes raw atomic positions and elemental types to construct and operate on tensors that encode the local chemical environment. A defining characteristic of this approach is the incorporation of rotational equivariance. This is achieved through network layers, often employing message passing or interaction blocks, that are specifically designed to manipulate these tensorial features in a manner consistent with their transformation properties under rotation of the coordinate system. For example, vector quantities representing interatomic orientations will rotate precisely as the physical system rotates. This architectural design ensures that derived vector properties, most notably atomic forces, are inherently equivariant, while scalar properties like the potential energy are inherently invariant. Such satisfaction of fundamental physical symmetries by construction, rather than solely through the pre-design of input descriptors, is thought to facilitate a more robust and fundamental learning of the complex relationship between atomic structure and energy. Finally, recent work in the TensorNet architecture allows the inclusion of different molecular attributes besides atomic coordinates to the input representation, opening the possibility of a more complex description of the atomic environment.[101]

2.4. Alchemical Free Energy calculations

Molecular Dynamics (MD) provides an atomistic, time-resolved view of molecular motion. In principle, a sufficiently long MD trajectory will sample the conformational space of a system,

allowing for the direct calculation of thermodynamic observables. However, many biological processes of interest, particularly protein-ligand binding and unbinding, involve transitions that are typically rare on the timescale of conventional MD simulations. For instance, the physical process of a ligand binding to its receptor involves the system navigating a complex path from the unbound state (two solvated molecules) to the bound state (a single complex). The direct, "brute-force" simulation of this process to equilibrium is often computationally intractable.[84,102]

Alchemical free energy calculations offer an alternative strategy to determine free energy (ΔG in a NPT ensemble) between two well-defined thermodynamic states of a system. These methods employ a non-physical, or "alchemical," pathway to connect the initial and final states. In these techniques, the potential energy function of a system, U, is gradually modified from that of an initial physical state (state 0, e.g., protein bound to ligand A) to that of a final physical state (state 1, e.g., protein bound to ligand B, or ligand A in solvent). This transformation is governed by a coupling parameter, denoted as λ, which varies smoothly from 0 (representing state 0) to 1 (representing state 1). The entire alchemical transformation is typically discretized into a series of intermediate, non-physical states, each corresponding to a specific fixed value of λ. Separate MD simulations are then performed for each of these λ-windows (or λ-states).[103] The energy is calculated with hybrid potentials that depend on λ. For example, linear interpolations use functional forms as:

$$V\left(r_N;\lambda\right) = (1 - \lambda)V_A\left(r_N\right) + \lambda V_B\left(r_N\right) \#(6)$$

In practice, more complex, non-linear mixing functions are often employed to improve the stability and convergence of the calculations, particularly when dealing with the creation or annihilation of atoms. A major numerical challenge, often called "endpoint catastrophe", in alchemical calculations arises when atoms are being created or annihilated during the

intermediate λ-windows. Near the endpoints of the transformation, where an atom's interactions are close to zero, other atoms can drift into the same physical space. When the interactions are turned on even slightly in the next lambda-window, this overlap leads to nearly infinite repulsive energies and forces, causing numerical instabilities that can crash the simulation. The standard solution to this problem is the use of "soft-core potentials".[104] These are modified versions of the Lennard-Jones and electrostatic potential functions that remain finite even at zero interatomic distance, but only for the intermediate, non-physical lambda states. This softening of the potential core prevents the endpoint catastrophe and allows for stable and convergent simulations. For instance, the LJ potential for disappearing atoms in intermediate states is :

$$V = 4\epsilon(1-\lambda)\left[\frac{1}{\left[\alpha\lambda+\left(\frac{r_{ij}}{\sigma}\right)^2\right]^2} - \frac{1}{\alpha\lambda+\left(\frac{r_{ij}}{\sigma}\right)^2}\right] \#(7)$$

These techniques are extensively used in fields such as drug discovery for the prediction of protein-ligand binding affinities, solvation free energies, and the effects of mutations on protein stability. Generally, these transformations involve changes in molecular topology (that defines the chemical identity and connectivity of the atoms that define a molecule's structure), such as mutating one ligand into another. Two main strategies for defining the system are employed: single-topology and dual-topology. In a single-topology approach, a hybrid molecular structure is created that contains all atoms necessary to represent both state A and state B. Atoms unique to state A are "annihilated" (their interactions are turned off) as λ goes from 0 to 1, while atoms unique to state B are "created" (their interactions are turned on). In a dual-topology approach, both molecule A and molecule B are present in the simulation system throughout the

transformation. However, at λ=0, only molecule A interacts with the environment (and itself, if it's being solvated or bound), while molecule B is present as a set of non-interacting "dummy" atoms.[105]

For the prediction of protein-ligand binding affinities, alchemical free energy calculations allow the estimation of absolute binding free energy (ABFE) and relative binding free energy (RBFE).

In the ABFE case, the ligand is transformed into a noninteracting state through the lambda windows, and special restraints are placed for the ligand to remain bound to the pocket as the interactions disappear. In contrast, the RBFE calculations estimate the difference between free energies of two ligands. ($\Delta\Delta G_{A-B}$). To do so, two independent simulations are performed transforming ligand A into ligand B in solvent and in complex, closing the thermodynamic cycle shown in Figure 8.[103]

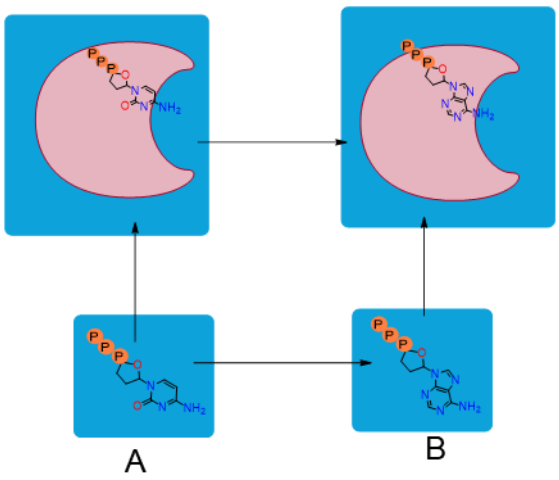


Figure 9 The thermodynamic cycle for relative binding free energy calculations. ligand A bound to protein (top left), ligand B bound to protein (top right), ligand A in solution (bottom left), and ligand B in solution (bottom right). Alchemical transformations (red arrows) convert ligand A to ligand B in both the protein-bound state (ΔG_{comp}) and in solution (ΔG_{solv}). The vertical arrows represent the physical binding processes for ligands A (ΔG_A) and B (ΔG_B).

The final $\Delta\Delta G_{A-B}$ is then estimated using the expression:

$$\Delta\Delta G_{A-B} = \Delta G_B - \Delta G_A = \Delta G_{complex} - \Delta G_{solv} \# (8)$$

To estimate $\Delta G_{complex}$ and ΔG_{solv}, the Thermodynamic Integration (TI) approach calculates the following integral:

$$\Delta G = \int_{\lambda} \langle \frac{\partial V(r_N,\lambda)}{\partial \lambda} \rangle d\lambda \# (9)$$

Because the simulations are performed at a series of discrete λ values. At each λ, the ensemble average $\langle \partial V(r\_N,\lambda)/\partial\lambda \rangle$ is computed from the MD trajectory and the hybrid potential. Then the integral is solved numerically via the trapezoidal rule, Simpson's rule, or more sophisticated Gaussian quadrature schemes[103] with a set of corresponding weights w_k as:

$$\Delta G \approx \sum_{k=1}^{M} w_k \langle \frac{\partial V(r_N,\lambda)}{\partial \lambda} \rangle \# (10)$$

2.5. Alchemical transfer method.

The Alchemical Transfer Method (ATM) is a novel approach that redefines the concept of an alchemical pathway. Instead of modifying the parameters of the potential energy function (e.g., atomic charges or LJ parameters), ATM performs the transformation through a coordinate displacement of the ligand. The entire process is conducted within a single simulation box, moving the ligand from a region in the bulk solvent to the receptor's binding site.

The alchemical parameter λ in ATM controls this spatial transfer (h), rather than directly scaling interaction potentials. In addition, the ATM method avoids end-point catastrophes by adding modifications to the potential energy as:

$$V_\lambda(r_N) = V(r_N) + W_\lambda(u) \#(11)$$

Where u is the perturbation energy:

$$u = U_{\lambda=1} - U_{\lambda=0} \#(12)$$

And W_λ is the softcore potential energy

$$W_\lambda(u) = \frac{\lambda_2 - \lambda_1}{\alpha} \ln ln\left[1 + e^{-\alpha\left(u_{sc}(u) - u_0\right)}\right] + \lambda_2 u_{sc}(u) + w_0 \quad \#(13)$$

Where the parameters λ_2, λ_1, α, u_0, and w_0 are functions of the current value of λ during the alchemical transformation.[106]

In the case of Relative binding free energy calculations, the ATM implementation involves a thermodynamic cycle of three components.[107,108] A starting point that contain the ligand A interacting with the protein and the ligand B interacting with the bulk (described by the potential U_{RA+B}), and the inverted cases, described by U_{RB+A}. Finally, a third intermediate state in which both ligands are simultaneously interacting at 50% strength in the binding site and solvent bulk $U_{\frac{1}{2}}$ with energy:

$$U_{\frac{1}{2}} = \frac{U_{RA+B} + U_{RB+A}}{2} \quad \#(13)$$

The final ΔG_{A-B} is then calculated as :

$$\Delta G_{A-B} = \Delta G_B - \Delta G_A = \Delta G_{leg_1} - \Delta G_{leg_2} \quad \#(14)$$

Where ΔG_{leg_1} involves the transformation from the state U_{RA+B} to $U_{\frac{1}{2}}$, and ΔG_{leg_2} involves the transformation from the state U_{RB+A} to $U_{\frac{1}{2}}$.

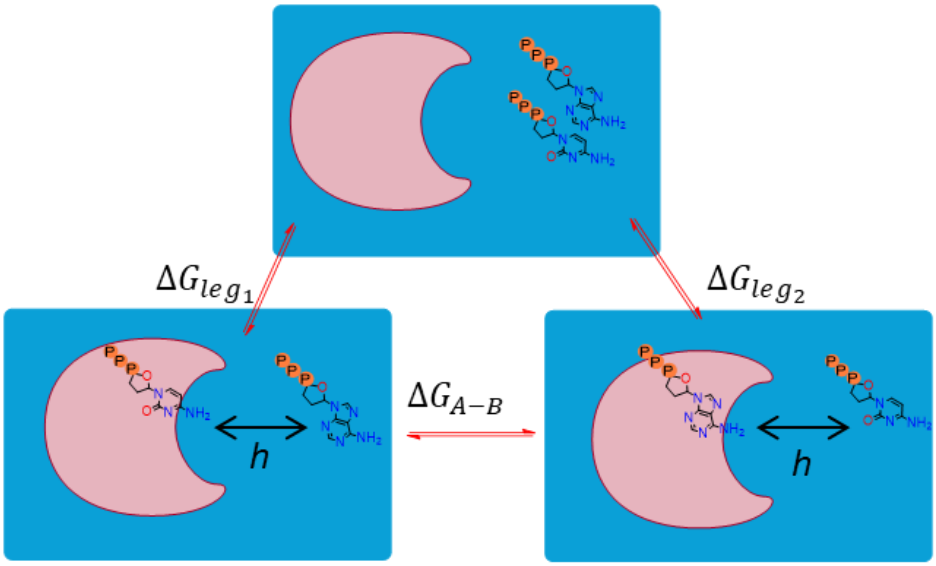


Figure 10 The thermodynamic cycle used in the ATM approach. (top) both ligands A and B in bulk solvent away from the protein; (bottom left) ligand A bound to the receptor with ligand B in bulk; (bottom right) ligand B bound to the receptor with ligand A in bulk. The alchemical parameter controls the spatial transfer (h) of ligands between the binding site and bulk solvent rather than modifying potential energy parameters.

For the ATM implementation, the perturbation energies collected at the different λ-states (for each leg of the pathway) can be effectively processed using the Unbinned Weighted Histogram Analysis Method (UWHAM),[109] which calculates the free energy of a certain λ-state by solving in a self-consistent manner the equation :

$$G_{\lambda} = -\ln ln\, Z_{\lambda} \quad \#(15)$$

$$Z_{\lambda} = \sum_{i=1}^{M} \frac{e^{u_{\lambda i}}}{\sum_{\lambda s} \sum_{j} N_j e^{G_{\lambda}} e^{u_{ji}}} \quad \#(16)$$

And the total free energy differences from the whole transformation $\lambda = 0$ to $\lambda = 1$ is calculated as:

$$\Delta G = \sum_{k} G_k \quad \#(17)$$

2.6. NNP applications in Free Energy calculations

NNPs are integrated into FEP workflows primarily through hybrid NNP/MM schemes. In this paradigm, the region of the system undergoing significant chemical or structural change, or requiring the highest accuracy (e.g., a ligand and its immediate protein binding site environment), is treated with an NNP trained on quantum mechanical data. The rest of the system, typically bulk solvent or distant protein domains, is modeled using a computationally less demanding classical MM force field. This hybrid approach aims to capture the detailed energetics of intramolecular interactions, such as ligand strain, and crucial intermolecular interactions within the NNP region with an accuracy approaching that of QM methods, while maintaining computational feasibility for the entire system. Such NNP/MM setups endeavor to bridge the traditional gap between the limited accuracy of purely classical FEP and the prohibitive computational cost of extensive QM/MM FEP.[42]

Specifically, the total potential energy (V_{total}) of such a hybrid system is generally formulated as a sum of three components:

$$V_{total} = V_{NNP(RI)} + V_{MM(RII)} + V_{NNP-MM(RI,RII)} \quad \#(18)$$

Here, RI represents the coordinates of atoms within the NNP region, and RII denotes the coordinates of atoms in the MM region. The $V_{NNP(RI)}$ term is the potential energy in the NNP region. Here, the NNP captures complex many-body effects, conformational energies, and intramolecular strain with an accuracy approaching that of the reference QM method. The $V_{MM(RII)}$ term corresponds to the potential energy of interactions occurring exclusively *in* the MM region, computed using standard classical force field functional forms (e.g., bonded terms, van der Waals, and electrostatic interactions). Finally, the $V_{NNP-MM(RI,RII)}$ term describes the coupling and interactions between the NNP region and the MM region. For practical and

efficient NNP/MM implementations, these cross-interactions, primarily non-bonded van der Waals and electrostatic forces, are commonly handled using parameters from the classical MM force field. This means that atoms in the NNP region are assigned MM-compatible parameters (e.g., partial charges, Lennard-Jones parameters) specifically for their interactions with the MM region atoms.[43]

This reliance on classical MM parameters for NNP-MM interactions is characteristic of a "mechanical embedding" scheme, or a simplified form of electrostatic embedding. In this approach, the NNP energy of the NNP region and the MM energy of the MM region (including their MM-described interactions) are essentially additive. More sophisticated "electrostatic embedding" schemes, in contrast, should aim for a more quantum-mechanically sound coupling where the NNP (or QM method in QM/MM) explicitly experiences the electrostatic field of the MM region, and its electronic structure (and thus energy and forces) responds dynamically to this field.[110]

2.7. Deep learning fundamentals.

Machine learning (ML) is a field of computer science that enables systems to learn from data and improve their performance on specific tasks without being explicitly programmed for each instance. Instead of relying on predefined rules, ML algorithms identify patterns and relationships within datasets to build models capable of making predictions or decisions. This approach has become integral to many modern applications, from computer vision, advancing scientific discovery, and specifically drug design. The core idea is to develop programs that can automatically adapt and enhance their capabilities by processing and analyzing new information, leading to systems that can tackle complex problems by learning directly from experience.[111]

Supervised learning is a dominant paradigm in ML where algorithms learn from datasets made of input features and their corresponding known output labels, guiding the learning process to infer a mapping function for predicting outputs for new, unseen inputs. This paradigm encompasses two main task types: regression and classification, distinguished by the nature of the outcome variable. Regression tasks involve predicting a continuous numerical output, such as binding affinity values. Performance is assessed using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared. Classification tasks, conversely, predict discrete, categorical labels, assigning inputs to predefined classes.[111]

A central challenge in the development of predictive models is the management of the bias-variance tradeoff. This tradeoff describes the inherent tension between a model's ability to accurately capture the underlying relationships in the training data (low bias) and its capacity to generalize these learned patterns to new, unseen data (low variance). Bias refers to the error introduced by approximating a potentially complex real-world phenomenon with a model that is too simplistic, leading to poor performance on both training and test datasets. For instance, a model predicting a complex non-linear relationship might have high bias if it is restricted to a simple linear form. Conversely, variance quantifies a model's sensitivity to small fluctuations or noise within the specific training dataset used. A model with high variance, perhaps a very complex model trained on a small dataset, might overfit by essentially memorizing the training data, including its inherent noise, leading to poor generalization to new data instances. The relationship is typically inverse: increasing model complexity (e.g., adding more features or parameters to a model) reduces bias but increases variance, and vice versa. The objective is to find an optimal complexity that minimizes total error, which comprises the square of the bias, the

variance, and an irreducible error inherent in the data itself. Techniques such as dimensionality reduction, feature selection, regularization, and cross-validation are crucial for managing this tradeoff and are contingent upon the specific problem and dataset characteristics.[112]

2.8. Principles of Deep Learning and Neural Network Training

Deep Learning (DL) is a subfield of machine learning characterized by Artificial Neural Networks (NNs) with multiple layers (deep architectures) designed to learn hierarchical representations of data, automatically discovering intricate patterns from raw input like images, text, or sensor readings. DL has achieved state-of-the-art results in complex tasks within drug discovery and bioinformatics, such as protein structure prediction, variant effect prediction, de novo drug design, and predicting protein-nucleic acid interactions.[113]

The core components of a NN include neurons (computational units), organized into one or more hidden layers and an output layer (predicting, e.g., a class label or a numerical value). Weights and biases are learnable parameters modulating connections and neuron offsets, respectively. In addition, these computational units have activation functions (e.g., sigmoid, tanh, ReLU) to introduce non-linearity, crucial for learning complex mappings from the input data. The training of most neural networks relies on the backpropagation algorithm, an efficient method for computing the gradient of the network's loss function (e.g., the difference between predicted and true values) with respect to all its learnable parameters. It involves a forward pass, where input data propagates through the network to compute outputs and loss, and a backward pass, where the error gradient is propagated backward using the chain rule of calculus to compute gradients for each layer's parameters. These gradients are then used by an optimization algorithm (e.g., SGD, Adam) to update parameters iteratively to minimize loss.[112,114]

During training, one of the most common problems in Deep learning models is overfitting, which occurs when a model learns the training data too comprehensively, capturing not only the underlying patterns but also the noise of the training set, which leads to poor performance on data not encountered during training. Regularization techniques are designed to prevent overfitting and enhance the generalization capability of a model, ensuring it performs well on new, unseen data. By controlling model complexity, it discourages overly intricate models that might fit the training noise rather than the true data distribution. Early stopping is a widely adopted and computationally inexpensive regularization technique. It involves monitoring the model's performance on a separate validation dataset (data not used for updating the model's parameters). As training progresses, while performance on the training set typically continues to improve, performance on the validation set often improves for a period and then may stagnate or begin to degrade. This divergence point usually signals the onset of overfitting[112]. Early stopping halts the training process when the validation performance no longer improves or starts to decline, based on a predefined patience criterion. The model parameters from the training epoch that yielded the best performance on the validation set are then typically saved and used as the final trained model. This method acts as an implicit regularizer by restricting the optimization process to a region in the parameter space that demonstrates better generalization. It is important to note, however, that using the validation set for both model selection and for estimating the final model performance can lead to an overestimation of the model's true generalization ability, therefore, a separate, untouched test set remains indispensable for a rigorous final evaluation. Other common regularization methods include L1 and L2 regularization, which add penalties to the loss function based on the magnitude of the model weights, and dropout, which randomly deactivates a fraction of neurons during training to prevent overfitting.[115]

2.9. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are specialized deep neural networks highly effective for processing data with a grid-like topology, making them invaluable in computational drug design and protein dynamics. Applications include analyzing 3D structures of proteins and ligands, processing molecular dynamics (MD) simulation trajectories, which represent protein movement over time, or interpreting volumetric data from cryo-electron microscopy (cryo-EM) maps.[113] Their architecture incorporates local receptive fields, where neurons connect only to small, localized regions of the input (e.g., a patch of atoms in a protein binding site or a short segment of an MD trajectory), and parameter (weight) sharing, where the same filter (kernel) is applied across different spatial or temporal locations. This design drastically reduces parameters compared to fully connected networks, making CNNs computationally efficient and less prone to overfitting when dealing with high-dimensional biomolecular data, while also providing translation equivariance (e.g., a specific chemical motif or protein secondary structure element is detected regardless of its absolute position). CNNs learn features hierarchically: early layers might detect low-level features while deeper layers combine these to recognize more complex patterns.[112]

A typical CNN architecture comprises a sequence of layers, usually starting with convolutional layers for feature extraction. These are often interleaved with activation layers (e.g., ReLU) for non-linearity, and pooling layers for downsampling and dimensionality reduction, helping to focus on the most salient features. These are followed by one or more fully connected layers for classification (e.g., identifying the object in an image) or regression (e.g., estimating a continuous value from the input) based on the extracted high-level features, and an output layer. Convolutional layers apply learnable filters (kernels) to the input. Each filter, typically a small

square matrix of weights (e.g., 3x3 or 5x5), slides across the input dimensions (height and width), computing the sum of the element-wise (or stepwise) multiplications between the filter and the input patch at each position. This process highlights the presence of the specific feature that the filter detects. The set of K filters in a layer produces K feature maps, forming the depth of the output volume. The dimensions of the output feature maps are determined by several hyperparameters: the size of the input volume (W, H), the filter size (F), the stride (S), and the amount of zero-padding (P) applied to the input borders (see Figure 10). Stride dictates how many pixels the filter moves at each step; a stride of 1 means the filter moves one pixel at a time, while a larger stride results in a smaller output. Padding involves adding zeros around the input volume, which can help control the output size and preserve information at the borders.[112] The output width (Wout) and height (Hout) are calculated as follows :

$$W_{out} = \left(\frac{W-F+2P}{S}\right) + 1$$
$$H_{out} = \left(\frac{H-F+2P}{S}\right) + 1$$

The depth of the output volume (D) is equal to the number of filters (K) used in the convolutional layer. These filters are learned during training via backpropagation.

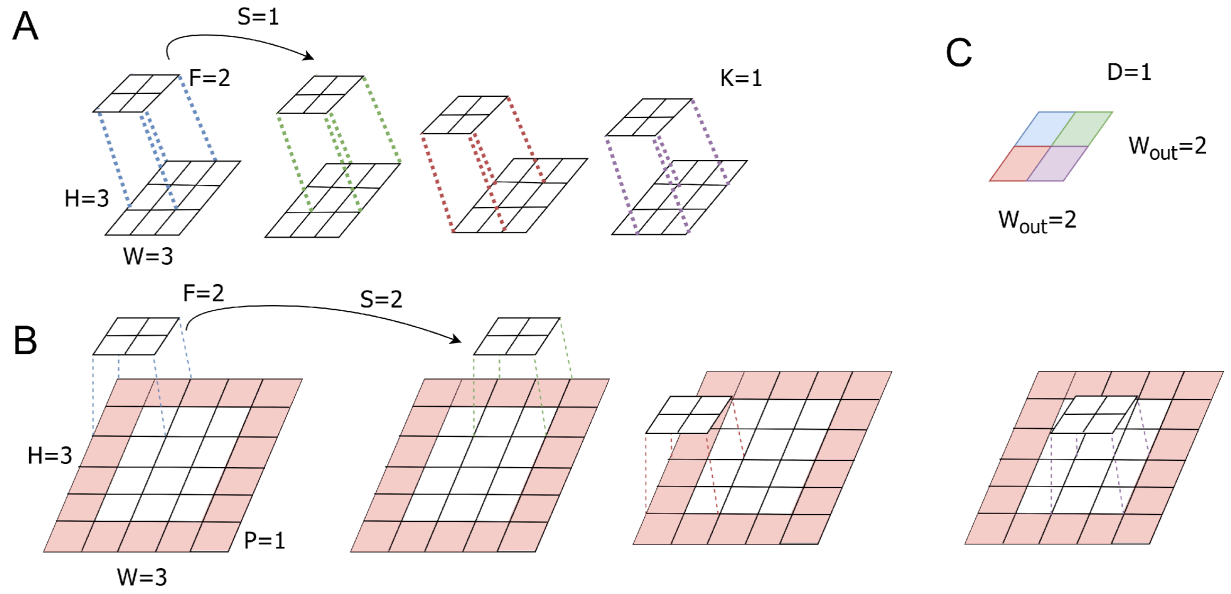


Figure 11 Schematic representation of the convolution operation. A) A Convolution of a 2x2 filter with a stride of 1. B) A convolution of a 2x2 filter with a stride of 2 and padding of 1. C) output map from both convolutions A and B.

Pooling layers (subsampling layers) are commonly used to reduce the spatial dimensions of feature maps, thereby decreasing computational load. Additionally, it creates representations robust to small translations or distortions in the input (contributing to translation invariance) and prevents overfitting. They operate by sliding a non-learnable filter and applying an aggregation function. The two most common types are Max Pooling, which selects the maximum value from the filter's region and is often preferred for preserving the most salient features (e.g., the strongest response to an edge detector), and Average Pooling, which calculates the average value, providing a smoother, more generalized representation of features within a region. Pooling layers act as an information bottleneck, selectively discarding some information; the choice of pooling type can impact model performance by influencing what information is preserved versus discarded.[112]

3. Correct nucleotide selection is confined at the binding site of polymerase enzymes

3.1. Introduction

As established in the preceding sections, the ability of polymerases to choose the correct nucleotide during DNA replication and maintenance is fundamental for numerous cellular processes.[1,2] Various mechanisms for nucleotide selection have been proposed. These have been widely explored in the context of Pol β, thanks to the extensive experimental and computational data available for this enzyme.[27,35,116–118] Specifically, for Pol β, it has been proposed that pre-chemistry steps are a critical part of the selection process. These steps are characterized by open-to-close conformational changes in the enzyme, driven primarily by the same transition in the finger N-subdomain.[22] Indeed, site-specific mutations, such as E295K, disturb the network of residue interactions responsible for these conformational shifts, resulting in significant alterations to the enzyme turnover number kpol, which measures the nucleotide incorporation rate.[119] Furthermore, mutations located near the active site have a direct impact on the protein during the chemical bond formation for dNTP incorporation,[120] and variations in local dynamics surrounding the metal center can also indicate crucial differences in fidelity. The energetic hurdle for creating the phosphodiester bond is frequently suggested to be the step that limits the overall rate.[10] Thus, any alterations in the stabilization of the transition state for nucleotide addition can have a profound effect on catalysis and, by extension, on fidelity.

Considering the wealth of kinetic and structural data available for Pol β, we aimed to dissect how nucleotide binding affects the selection of correct and incorrect bases, under a certain template (dG). Specifically, we studied the insertion of correct and incorrect nucleotides into the wild-type Pol and its R283A mutant, as experimentally investigated by Ahn et al.[29] Notably, the mutant R283A was found to reduce the binding affinity of both correct and incorrect nucleotides, affecting fidelity significantly. Experimental findings revealed that this particular mutant, despite its 14.8 Å distance from the catalytic site, impairs the capacity of Pol β to distinguish between correct and incorrect insertions. In detail, for the correct dG:dCTP match, the k_{pol} was measured at 9.4 s^{-1} for the WT and 0.83 s^{-1} for the R283A mutant. Additionally, a decrease was observed in the ratio of binding affinities between mismatched and matched pairs; the ratio of Kd(dG:dATP)/Kd(dG:dCTP) was 24 for the WT, but only 2.6 for the R283A mutant. This shift in kinetic and thermodynamic properties resulted in a dramatic change in fidelity, which fell from 1 error per ~14000 insertions for the WT to 1 error per ~150 insertions for the R283A mutant. However, the underlying mechanistic cause for this mutant's significant effect on nucleotide selectivity in Pol β was not understood.

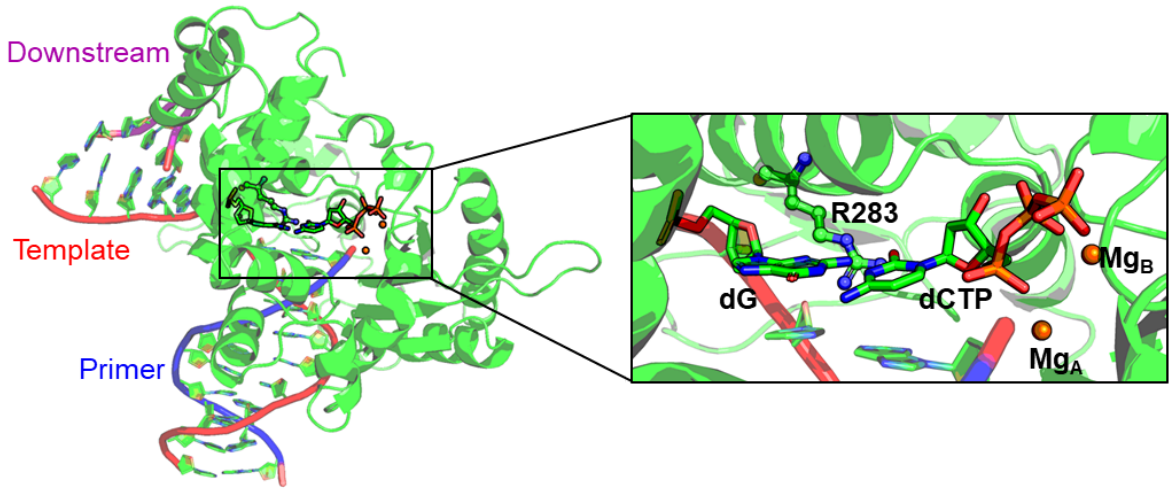


Figure 12. Overall structure of Pol β (green) bound to primer-template DNA showing the downstream DNA strand (magenta), template strand (red), and primer strand (blue). The incoming dCTP is positioned opposite the template dG at the active site. (Inset) Magnified view of the catalytic site showing residue R283 positioned 14.8 Å from the catalytic magnesium MgA, where it interacts with the template base dG. The two catalytic metal ions MgA and MgB (orange spheres) coordinate the primer terminus and dCTP triphosphate. Despite its distance from the catalytic center, R283 plays a critical role in nucleotide selectivity.

Our computational study, based on molecular dynamics simulations and free energy calculations, allowed us to pinpoint key structural and dynamic characteristics localized to the catalytic site that affect the proficiency of Pol β in discriminating between matched and mismatched base pairings. Structural alignment of polymerases in the X-family reveals that the conserved position of the residue R283A and a particular set of interactions centered around the nascent base pair are intimately linked to the choice between correct and incorrect nucleotides.

3.2. Results

3.2.1.Downstream DNA and Close state analysis in MD simulations

First, we analyzed two models of matched base pairing, in which dCTP forms a correct pair with the template dG, both with and without the inclusion of the downstream strand. In both instances,

these dG:dCTP·DNA·Pol β ternary systems (PDBID 4KLQ) demonstrated stability and remained in close proximity to their crystallographic conformations. The protein backbone RMSD was 1.9 ± 0.3 Å and 1.6 ± 0.3 Å when the downstream strand, originally present in the crystallographic structure ,was absent or present, respectively. Similarly, the systems modeling the mismatched dG:dATP base pair also exhibited a stable conformation, with a protein backbone RMSD of 2.1 ± 0.3 and 2.1 ± 0.3 Å, without and with the downstream strand, respectively. However, we observe that when the downstream strand was not present, the template DNA filament showed considerable fluctuation, with an RMSD of 4.6 ± 0.8 Å for the matched complex and 4.5 ± 0.6 Å for the mismatched one. These values can be contrasted with the models that included the downstream strand, which yielded RMSD values of 2.4 ± 0.5 Å and 3.3 ± 0.7 Å for matched and mismatched complexes, respectively (See Table 1). This comparison illustrates the stabilizing influence of the downstream strand on the template DNA filament.

Table 1. Average RMSD ± standard deviation for dG:dCTP-DNA-Pol β (matched), dG:dATP-DNA-Pol β (mismatched aligned catalytic center) and dG:dATP-DNA-Pol β (mismatched misaligned catalytic center).

System	label	WT NoD-DNA	R283A NoD-DNA	WT D-DNA	R283A D-DNA
dCTP:dG Pol β	Protein	1.95 ± 0.32	2.07 ± 0.37	1.57 ± 0.26	1.97 ± 0.38
	DNA	4.56 ± 0.83	5.43 ± 1.43	2.40 ± 0.50	4.91 ± 0.74
	dCTP	0.58 ± 0.24	0.71 ± 0.22	0.54 ± 0.22	0.81 ± 0.11
	N-Subdomain	1.83 ± 0.39	1.84 ± 0.44	1.28 ± 0.25	1.91 ± 0.40
	a Helix	1.65 ± 0.44	1.88 ± 0.56	1.18 ± 0.27	1.79 ± 0.46
dATP:dG (aligned) Pol β	Protein	1.60 ± 0.40	1.61 ± 0.21	1.72 ± 0.44	1.53 ± 0.22
	DNA	3.10 ± 0.74	2.67 ± 0.53	2.98 ± 0.46	2.41 ± 0.53
	dCTP	0.42 ± 0.12	0.44 ± 0.20	0.82 ± 0.13	0.40 ± 0.08
	N-Sub	1.62 ± 0.46	1.76 ± 0.38	1.73 ± 0.42	1.37 ± 0.32
	a Helix	1.52 ± 0.66	1.23 ± 0.29	1.38 ± 0.51	1.05 ± 0.33
dATP:dG (misaligned) Pol β	Protein	2.10 ± 0.27	2.77 ± 0.47	2.07 ± 0.26	2.34 ± 0.30
	DNA	4.47 ± 0.60	3.99 ± 0.38	3.31 ± 0.67	5.83 ± 1.14
	dATP	1.21 ± 0.29	0.48 ± 0.20	0.62 ± 0.15	0.91 ± 0.17
	N-Sub	1.93 ± 0.43	3.26 ± 0.55	1.72 ± 0.33	3.50 ± 0.64
	a Helix	1.74 ± 0.48	3.46 ± 0.66	1.84 ± 0.44	3.38 ± 0.54

The protein's conformation consistently adopted what is known as the "closed state" across all simulated systems. In this state, the positioning of residues surrounding the nascent base pair, including R256 and F272, facilitates the correct coordination of D192 with the catalytic MgA ion (see Figure 12).[17,121] This specific conformational state was preserved regardless of whether the downstream strand was present or absent. The integrity of this closed state was primarily defined by two structural characteristics: i) the stability of the α-helix located within the finger domain (with an RMSD of 1.6 ± 0.4 Å for matched and 1.7 ± 0.5 Å for mismatched systems), which did not deviate from its closed conformation, and ii) the substantial distance maintained between residues R258 and D192, which measured 10.9 ± 0.6 Å and 10.9 ± 0.7 Å for the matched and mismatched complexes, respectively (Table 2). These observations reaffirmed the general stability of these ternary complexes while in the closed state, which is consistent with our prior simulations of comparable ternary models of Pol β. Time series for the RMSD and close conformation distances are presented in Figure S1Figure S6.

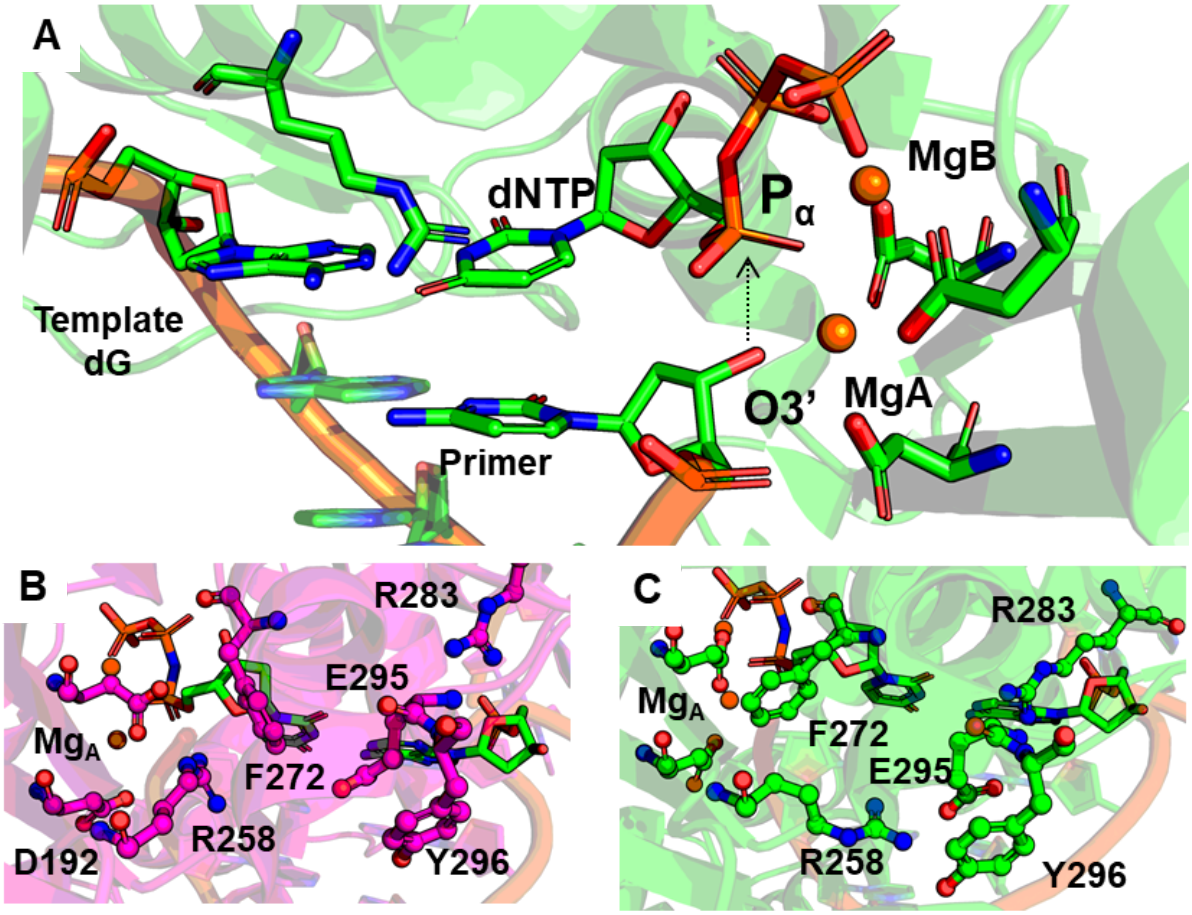


Figure 13 A) Closer view to the Pol B catalytic site. B-C) Close view of the interaction network for D192, close to the catalytic metal (MGA), in open (pink) and closed (green) conformations. Note the orientation of R258 toward D192 in the close conformation (C).

Table 2. Average distances ± s.d for characteristic close conformation in Pol β.

System	label	Ref	WT NoD-DNA	R283A NoD-DNA	WT D-DNA	R283A D-DNA
dCTP:dG Pol β	F272-D192	4.9	5.32 ± 0.34	6.96 ± 0.90	5.29 ± 0.23	7.03 ± 0.71
	F272-R258	6.1	6.60 ± 0.38	5.83 ± 0.44	6.47 ± 0.44	5.63 ± 0.44
	R258-D192	10.3	11.14 ± 0.41	11.15 ± 0.51	10.89 ± 0.63	10.93 ± 0.37
	R258-E295	5.3	5.13 ± 0.20	5.03 ± 0.25	5.12 ± 0.22	4.94 ± 0.23
	R258-Y296	6.7	7.44 ± 1.04	7.72 ± 0.91	6.65 ± 0.43	7.13 ± 0.79
	E295-R283	7.4	7.74 ± 0.30	11.31 ± 0.51	7.77 ± 0.40	11.47 ± 0.50
dATP:dG (aligned) Pol β	F272-D192	4.9	7.18 ± 0.84	7.62 ± 0.30	5.97 ± 1.16	7.41 ± 0.42
	F272-R258	6.1	5.58 ± 0.44	5.65 ± 0.47	4.97 ± 0.52	5.33 ± 0.42
	R258-D192	10.3	11.04 ± 0.46	11.45 ± 0.51	9.13 ± 0.78	10.95 ± 0.36
	R258-E295	5.3	5.46 ± 0.37	4.94 ± 0.27	5.05 ± 0.57	4.86 ± 0.19
	R258-Y296	6.7	6.51 ± 0.31	8.23 ± 0.79	8.16 ± 1.48	7.90 ± 0.66
	E295-R283	7.4	7.22 ± 0.62	11.37 ± 0.54	8.82 ± 1.09	11.45 ± 0.46
dATP:dG (misaligned) Pol β	F272-D192	4.9	7.98 ± 1.48	6.19 ± 0.97	6.26 ± 0.79	7.92 ± 0.44
	F272-R258	6.1	5.58 ± 0.42	6.16 ± 0.70	6.27 ± 0.74	5.33 ± 0.34
	R258-D192	10.3	9.56 ± 0.71	11.02 ± 0.87	10.90 ± 0.76	10.74 ± 0.56
	R258-E295	5.3	6.62 ± 0.97	4.62 ± 0.65	6.06 ± 0.72	5.60 ± 0.83
	R258-Y296	6.7	6.94 ± 1.09	7.05 ± 0.76	7.03 ± 0.77	7.27 ± 0.69
	E295-R283	7.4	9.20 ± 0.89	12.33 ± 0.76	9.64 ± 1.21	11.70 ± 0.77

Subsequently, we turned our attention to the active site, specifically focusing on the placement of the nucleophilic oxygen O3' on the primer strand, a crucial element for catalysis. In the simulations involving the matched ternary complex, Pol β consistently showed a well-aligned, pre-reactive geometry (see Figure 12A).[10] The O3' atom was correctly coordinated to MgA (at a distance of 2.2 ± 0.1 Å) and was positioned for a nucleophilic attack on the Pα of dCTP, indicated by an O3'-Pα-Oα angle of 151º ± 8º. It is noteworthy that the presence of the downstream DNA did not alter this active site geometry, with the MgA-O3' distance and O3'-Pα-Oα angle remaining at 2.2 ± 0.1 Å and 150.8º ± 7º, respectively. In contrast, the mismatched dG:dATP model, initiated from a misaligned structure, consistently maintained a non-reactive state. The nucleophilic O3' stayed uncoordinated from MgA and, over the course of the simulations, its distance from the Pα of dATP increased to 6.1 ± 1.5 Å. Furthermore, the O3'-Pα-Oα angle remained misaligned at 115º±18º (Table 3).

To further probe the tendency of the mismatched system to favor a misaligned state, we initiated a simulation where the mismatched pair started in an aligned catalytic arrangement. This starting configuration was generated through an alchemical transformation of the correct dG:dCTP pair. Even in this scenario, the system quickly deviated from its optimal geometry. The MgA-O3' coordination and the O3'-Pα distance drifted from their ideal starting values of ~2.5 Å and ~3.2 Å to 4.7 ± 0.4 Å and 5.9 ± 0.7 Å, respectively, in the absence of the downstream strand. Moreover, the critical O3`-Pα-Oα angle averaged around ~142º, deviating from the optimal angle. This structural breakdown of the catalytic site occurred relatively quickly (after ~250 ns) in the absence of downstream DNA but took significantly longer (~750 ns) when the downstream DNA was present, once again highlighting its stabilizing function. Collectively, these findings confirm that a matched base pair in Pol β's ternary complex promotes a stable, aligned catalytic center, whereas a mismatched pair results in a misaligned and non-reactive geometry. Time series for the stability of the catalytic distances are presented in Figure S7-S9.

Table 3: Geometrical parameters of the catalytic site in the studied systems.

System	label	Ref	WT NoD-DNA	R283A NoD-DNA	WT D-DNA	R283A D-DNA
dCTP:dG Pol β	dNTP(Pa)- DA3(O3)	3.6	3.34 ± 0.18	3.64 ± 0.29	3.32 ± 0.15	3.44 ± 0.24
	dNTP(Pa)-MgA	3.8	3.36 ± 0.13	3.40 ± 0.13	3.35 ± 0.11	3.43 ± 0.18
	MgA-DA3(O3)	2.5	2.25 ± 0.16	2.20 ± 0.18	2.27 ± 0.15	2.14 ± 0.09
	MgA-D190	2.8	2.90 ± 0.09	2.93 ± 0.09	2.90 ± 0.08	2.95 ± 0.08
	O3'-Pα-Oα	156.0	151.02 ± 8.23	142.90 ± 13.43	150.79 ± 7.44	156.24 ± 7.24
dATP:dG (aligned) Pol β	dNTP(Pa)- DA3(O3)	3.6	5.37 ± 1.17	3.83 ± 0.20	3.90 ± 0.82	3.53 ± 0.23
	dNTP(Pa)-MgA	3.8	4.30 ± 0.74	3.40 ± 0.08	3.68 ± 0.57	3.37 ± 0.17
	MgA-DA3(O3)	2.5	4.15 ± 1.10	2.19 ± 0.10	2.98 ± 1.07	2.14 ± 0.09
	MgA-D190	2.8	3.01 ± 0.09	2.96 ± 0.07	2.96 ± 0.09	2.97 ± 0.07
	O3'-Pα-Oα	156.0	142.02 ± 9.03	132.30 ± 11.67	141.65 ± 11.90	156.97 ± 5.28
dATP:dG (misaligned) Pol β	dNTP(Pa)- DA3(O3)	3.6	6.93 ± 1.91	6.64 ± 0.39	6.15 ± 1.55	6.62 ± 0.49
	dNTP(Pa)-MgA	3.8	5.31 ± 0.16	3.43 ± 0.06	5.03 ± 0.33	3.44 ± 0.06
	MgA-DA3(O3)	2.5	6.33 ± 1.35	5.77 ± 0.35	5.98 ± 1.90	5.80 ± 0.39
	MgA-D190	2.8	3.05 ± 0.07	2.96 ± 0.08	3.04 ± 0.08	2.95 ± 0.07
	O3'-Pα-Oα	156.0	115.67 ± 18.68	142.17 ± 7.91	136.20 ± 10.86	140.80 ± 7.34

3.2.2.R283A mutant shows similar dynamics for matched and mismatched base pairs.

To investigate this puzzling effect of R283A over nucleotide selection, we conducted a deeper analysis of our simulations of the WT systems for both matched and mismatched pairs. We observed that within the WT matched dG-dCTP complexes that maintained a stable and aligned catalytic center, the guanidine group of residue R283 formed consistent hydrogen bond interactions with the N2 atom of the template dG base. This hydrogen bonding was maintained for 90% and 80% of the simulation duration, respectively, for systems with and without the downstream DNA substrate (Table S1). In contrast, within the WT mismatched dG-dATP complexes, which lacked a reactive conformation due to reactant misalignment, the guanidine group of R283 did not form an interaction with the N2 atom of the template dG base. An exception was noted in the case of the mismatched complex (generated with an alchemical free energy calculation) that began with an aligned catalytic center , where the R283's guanidine group did sustain a stable interaction with the dG's N2 atom (evidenced by a 90% hydrogen bond occupancy, Table S1). These findings, therefore, imply that the hydrogen bond interactions between R283 and the N2 atom of the dG template base promote the proper alignment of reactants and enhance the overall stability of a reactive conformational state in Pol β.

Following previous mutagenesis experiments, when the R283 residue was mutated into alanine in our ternary complexes, we found that the ternary mismatched dG:dATP−DNA−Pol β model remained dynamically aligned at its catalytic center. This, again, occurs despite the mismatched dG:dATP pairing, which instead caused reactants misalignment in the WT system. In detail, we faced the following three cases: i) in the matched complex of the R283A mutant, the protein backbone was stable (RMSD 2.1 ± 0.4 Å), with an optimal alignment at the catalytic center, during the whole simulation, regardless of the presence/absence of the downstream DNA.

Specifically, the MgA-O3` distance was 2.2 ± 0.1 Å and the O3`-Pα-Oα showed a proper angle for nucleophilic attack (142º ± 5 º) (see Figure 13 A-D); ii) in the mismatched R283A model built from the mismatched WT X-ray structure PDB ID 4KLQ, the catalytic site remained misaligned with the distances O3`−Pα and MgA−O3` with values of 6.6 ± 0.5 Å and 5.8 ± 0.4 Å, respectively (see Figure S8). Removal of the downstream DNA did not change these structural features. Furthermore, throughout the simulations, the model maintained the initial close conformation. We confirmed this by monitoring key interactions involved in the open/close transition (Figure S5-S6), which maintained values characteristic of the closed state; iii) finally, the complex in which the mutant R283A started from a mismatched and aligned catalytic center maintained the initial geometry and overall structural features for catalysis. Namely, the MgA-O3` was 2.2 ± 0.1 Å. Similar geometrical features were observed also in the system containing the downstream DNA (Figure 13 A-D). This evidence from our simulations differs significantly from what was observed in the simulations of the WT mismatched aligned system, where the reactants gradually drifted away from the optimal catalytic position.

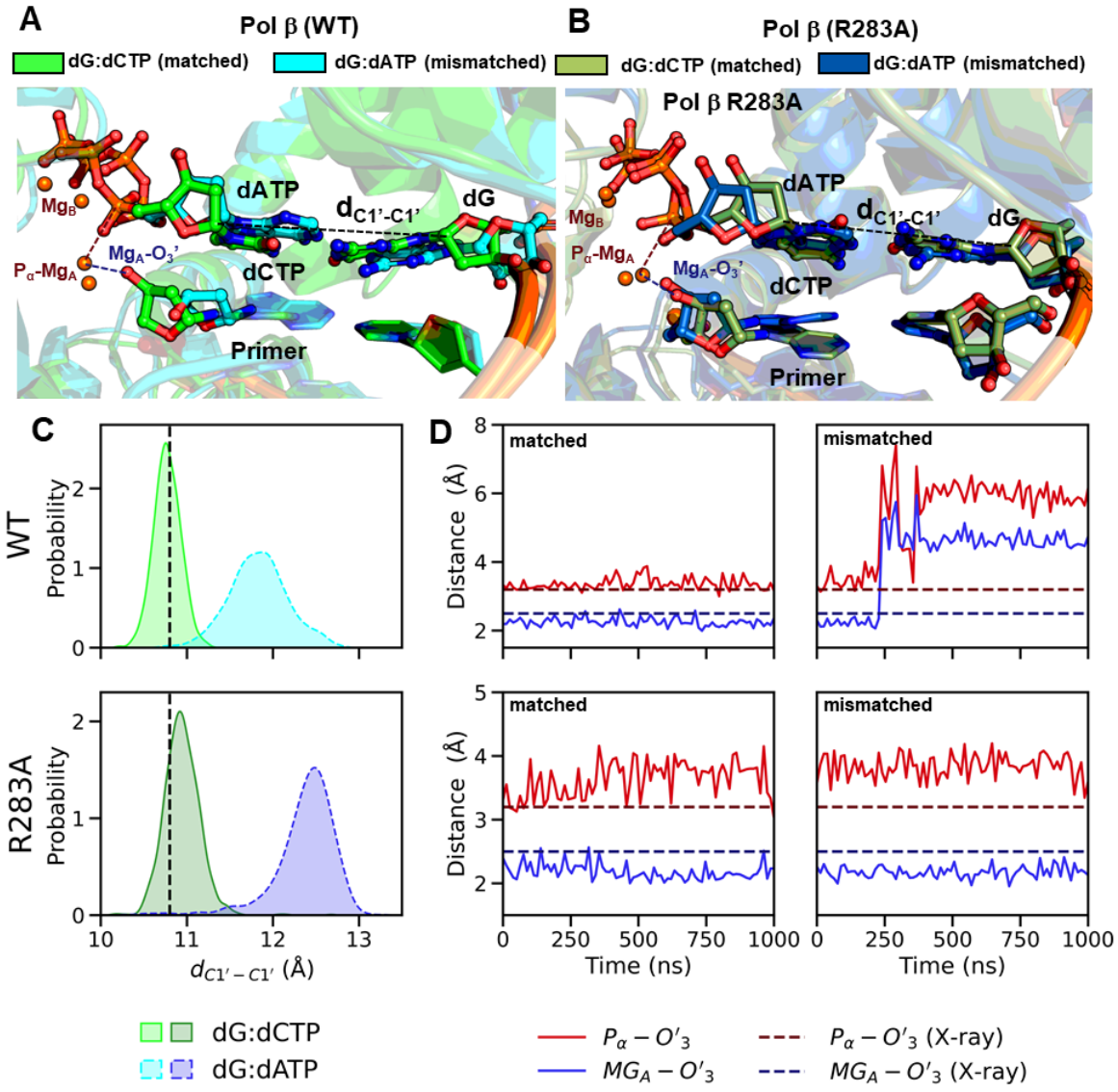


Figure 14 Comparison of matched (dG:dCTP in green tones) and mismatched (dG:dATP in blue tones) WT and R283A (A, B, respectively) complexes. O3′ is aligned for both matched and mismatched only in the R283A mutant. (C) Distribution of the distance C1′−C1′ (dC1′-C1′) in the nascent base pair. dC1′-C1′ reflects possible distortions on the helical parameters of the DNA upon mismatched complexes, as shown by the shifting of the distribution of the matched vs mismatched systems. The black dotted line represents the ideal dC1′-C1′ for a G:C base pair in B-DNA. (D) Stability of the catalytic center during equilibrium simulations, where the WT systems show the correct arrangement of the catalytic moieties only for matched systems. On the other hand, the R283A systems showed a stable catalytic site in both matched and mismatched complexes. Similar geometrical parameters were observed in models with the downstream DNA strand.

Intriguingly, within the R283A mutant systems, we detected a dynamic reorganization of the residues surrounding the nascent base pair. This primarily influenced the capacity of residue Y271 to directly interact with the O3' of the primer. Previous studies have highlighted this residue as being significant for catalysis, showing that its mutation to alanine compromised Pol β's ability to distinguish between right and wrong incoming nucleotides.[122] In our equilibrium simulations of the WT systems, Y271 only moves nearer to the N3 atom of the template dG in mismatched complexes (with dGN3-Y271OH distances of 3.7 ± 1.4 Å for dG:dATP-WT and 6.2 ± 0.4 Å for dG:dCTP-WT, see Figure 14 A-B). This positioning suggests a non-coplanar base pair arrangement in the case of mismatched complexes, which in turn appears to be a critical interaction for the ability to differentiate correct from incorrect nucleotides in the WT systems. In contrast, the R283A mutation introduces additional volume into the binding site due to the lack of the R283 side chain, resulting in a conformational shift of Y271, and therefore, occupying a different position in the binding site compared to its location in the WT systems. Consequently, the dGN3-Y271OH distance in both matched and mismatched complexes of the mutated enzyme reached comparable values of 5.2 ± 0.7 Å and 5.3 ± 0.4 Å, respectively. This means that in both mutated systems (the matched dG:dCTP-R283A and the mismatched dG:dATP-R283A, see Figure 14), the R283A mutation fosters a structurally very similar binding site. This evidence provides a mechanistic explanation for the R283A mutant's diminished capacity to distinguish between right and wrong base pairings at the catalytic site when contrasted with the WT systems.

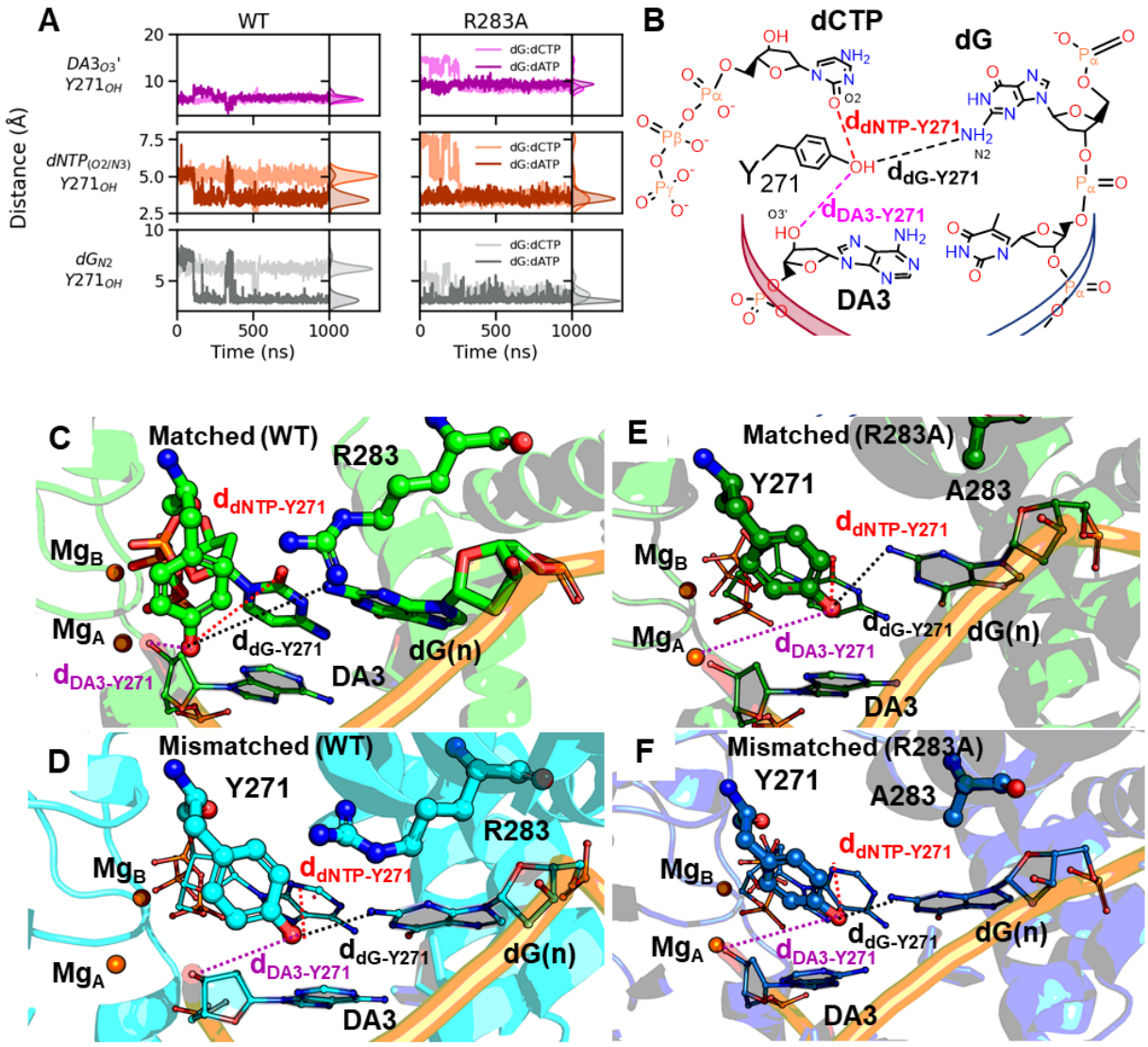


Figure 15 (A) Time series of interatomic interactions of Y271, which sits close to the catalytic site, with the incoming nucleotide (dNTP), the primer (DA3) without a downstream strand in the DNA substrate. (B) 2D representation of interatomic interactions of Y271 in the Pol β binding site. (C−F) Representative snapshots of matched and mismatched WTs (C−D) and R283A (E−F) complexes. These snapshots show the relative position of the primer DA3 and dNTP to the residues R283 and Y271. Interruption of planarity of base pairs by Y271 relative position is captured by the purple distance dY271−DA3. Nucleotides are represented in thinner sticks, and residues of the protein are shown as balls and sticks.

3.2.3.RBFE reproduce experimental values of binding affinities.

To compare the binding of mismatched (dATP) and matched (dCTP) nucleotides against a dG template, we calculated their relative binding free energy for both WT and R283A Pol β systems, with and without the downstream DNA. We used an alchemical transformation to convert dCTP into dATP in both solution and protein environments, treating the nucleobases with softcore potentials to decouple cytosine while coupling adenine (see Computational details). Throughout these calculations, the catalytic site and the enzyme's closed conformation remained stable (Figure S11. Catalytic distances for all replicas during the alchemical free energy calculation for the system Pol β–No DownstreamFigure S11-S14). Additionally convergence was verified in Figure S 15)

The ΔΔG for the C→A transformation within an aligned catalytic site conformation for the WT was found to be 7.7 ± 1.1 and 6.6 ± 0.5 kcal mol-1, with and without the downstream strand, respectively. These calculations effectively convert a matched complex into a mismatched one. The energetic cost for the same transformations in the R283A mutant was significantly lower, at 1.7 ± 1.1 and 0.9 ± 0.6 kcal mol-1, with and without the downstream strand, respectively. It is important to recognize that the energy for a C→A transformation inherently includes the cost of converting the more stable misaligned catalytic site of a mismatched complex into its less favorable aligned conformation. The interconversion of misaligned to aligned mismatched Pol β systems was previously calculated, using a combination of umbrella sampling and REST2 enhanced sampling simulations, with a total of approximately 3 kcal mol-1 in Pol β. By factoring in this ~3 kcal mol-1 difference between misaligned and aligned local conformations in our mismatched systems,[10] our calculated relative free energy of binding for dATP versus dCTP (template dG) in the WT is adjusted to ΔΔG = 4.7 ± 1.1 kcal mol-1 for the model with the

downstream strand and 3.6 ± 0.5 kcal mol-1 for the model without it. Both of these estimates are about 2 kcal mol-1 higher than the experimental values of 2.8 and 1.9 kcal mol-1 (See Figure S16). Despite this, our computational results accurately replicate the experimental trend and capture the reduced ability of the R283A mutant to differentiate nucleobases in Pol β. This reasonable agreement with the experimental trend, in turn, validates our models and mechanistic interpretations. Moreover, these findings show that alchemical free energy calculations can serve as a qualitative tool to estimate the relative binding free energy between matched and mismatched base pairs in the ternary complexes of Polymerases, and thereby to probe the impact of mutations on dNTP binding.

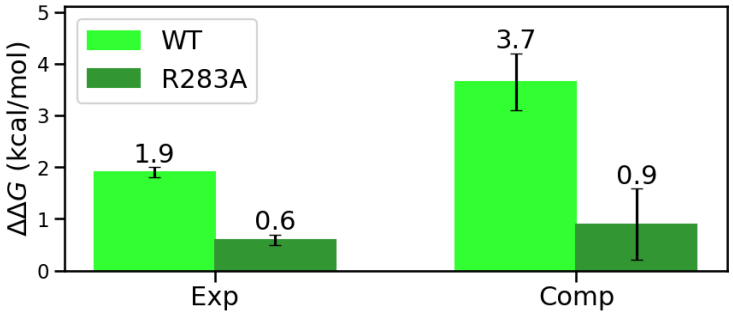


Figure 16. Results of alchemical free-energy calculations, obtained as an average of 10 replicas. Each value of ΔΔG calculated for WT considers the energetic cost of going from a reactive to a nonreactive catalytic site, based on the misalignment of the O3′ calculated previously as 3 kcal mol$^{-1.}$ Experimental values were calculated using (ΔG = RT ln(K)). Results on systems without downstream DNA.

3.2.4.Structural alignment of Polymerase family X and the role of a conserved R283 residue.

Finally, we performed residue conservation analyses of Pol β using Consurf.[123] We used 150 sequences with at least > 65% sequence identity. The analysis revealed that residues Y271, N279 and R283 (residue numbers are for Pol β), located around the nascent base pair, are conserved with a frequency of 86, 93 and 94%, respectively (Figure S 17 Conservation percentages of

discussed residues of MSA performed with consurf.). Y271 interacts with the primer nucleophilic end in the experimental structure and our simulations confirmed a stable interaction. Additionally, N279 maintained a close interaction during the simulation with the nucleobase of the incoming nucleotide as displayed in the experimental structure. Indeed, previous mutagenesis studies demonstrated the functional role of these two residues. Variants of Y271 and N279 display a change in the binding (Km) of the incoming nucleotide or its reactivity (kcat). Interestingly, Y271 is considered a steric gate, which contributes to the discrimination against ribonucleotides.[124] Then, we compared the structural data of polymerases from the X family, further confirming that R283 displays similar positioning and orientation around the template nucleotide in different snapshots of diverse polymerases such as β, λ, μ and in DNA nucleotidylexotransferase (TDT, See Figure 16). This evidence suggests that the mechanistic role of R283 in relation to fidelity in Pol β may be extended also to other members of the X family.

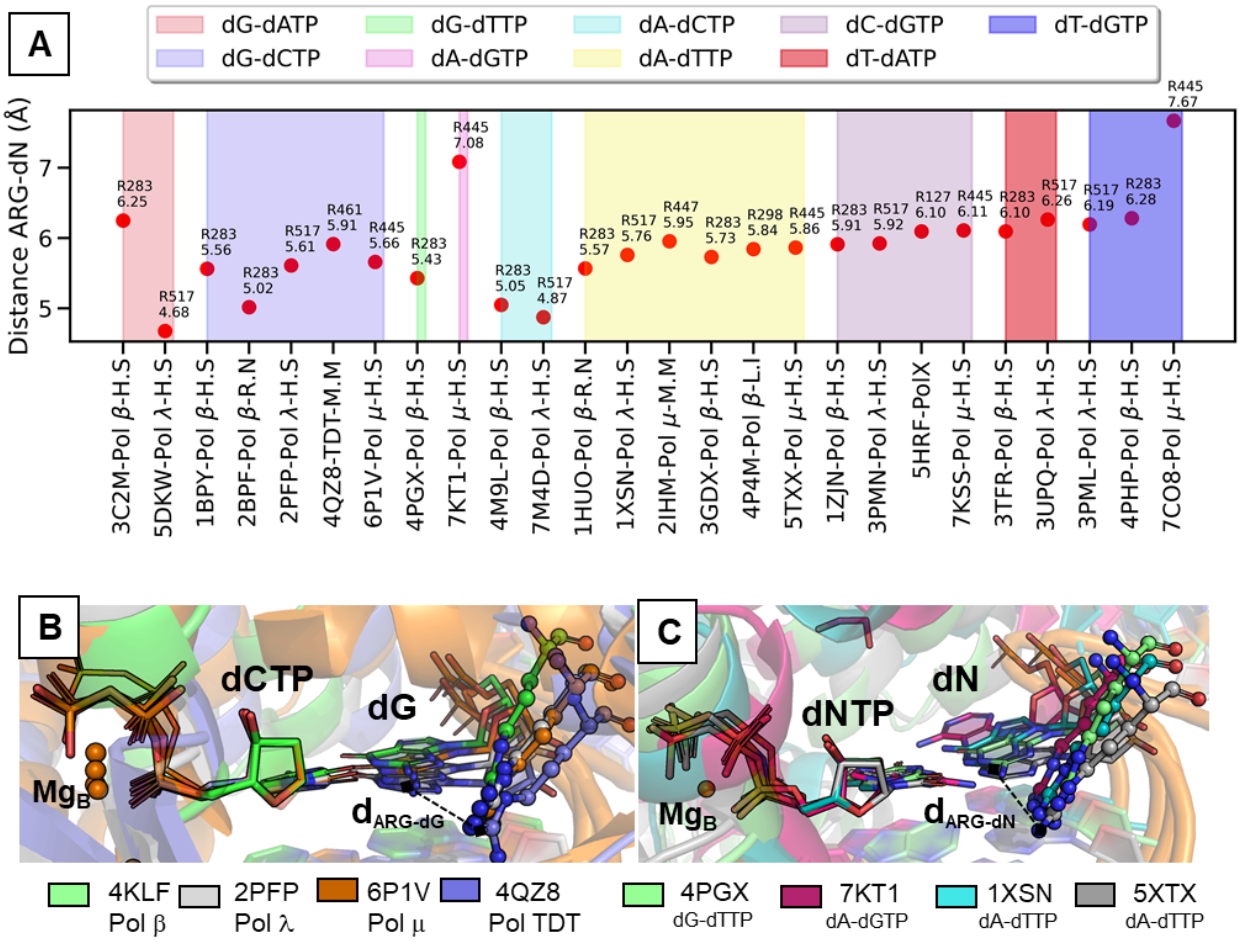


Figure 17. A) Distance of the center-of-mass of R283 from the template base (dG) in Pol β and of the equivalent arginine residue in other polymerases in the X-family, from different organisms (Homo sapiens H.S, Mus musculus, M.M, Rattus norvegicus, R.N, Leishmania infantum, L.I, and African swine fever virus, Pol X). These values cover both matched and mismatched cases. B) Structural alignment for matched dG:dCTP showing the nucleotides in thinner sticks. The arginine is showed in balls and sticks. C) Mismatched complexes in different polymerases show that the template dN and the arginine residue (R283 in Pol β) conserve a similar distance.

3.3. Computational details

Model systems. Pol β models were build based on the available crystallographic data: the dG:dCTP ternary complex was build from the structure 4KLF (1.85 Å resolution), while the mismatch dG:dATP was build from 4KLQ. Protonation states were assigned assuming pH 7.4 using the pdb2pqr software.[125] Complexes were solvated with TIP3P waters in a rhombic dodecahedral simulation cell, with additional ions K+,Cl- included to match the ionic

concentration used in kinetic experiments (2.5 mM MgCl2, 500 mM KCl) We considered models both including and non including the downstream strand (D-DNA) used in the crystallized structures. The template and primer strands of DNA were modified using the W3DNA[126] server to match the sequence used in kinetic experiments. (Template strand 3'-TTGGTTGAGTXCGAGC-5' where X is the template base G.). The R283A mutation was introduced using PyMol.[127]

Molecular dynamics. The protein and DNA were described using AMBERFF14SB[86] and parmbsc1[128] force fields, respectively. For Mg+2 ions, Van der Waals parameters proposed by Allner et al were used.[129] Charges for dCTP and dATP molecules were taken from the REDDB database.[130] Models were energy minimized for 2500 steps with a restraint of 300 kcal/mol on the protein backbone and dCTP atoms, followed by 25000 steps without restraints. Then, we performed NVT molecular dynamic simulation at 100 K for 100 ps. The heating was performed, including restraint of 300 kcal/mol on protein and DNA backbone atoms, using NPT MD at 100, 200 and 300 K for 100 ,300 and 500 ps respectively, followed by an additional ns without restraints. Afterward, we performed 1 μs of equilibrium MD, using Andersen thermostat and Monte Carlo barostat to control temperature (300K) and pressure (1 atm). Systems of mismatched ternary complexes in aligned catalytic conformation were simulated for 1 μs equilibrium MD starting from the last frame of the corresponding alchemical transformation (see above).

Alchemical binding free energy. We performed lambda-hopping alchemical calculations to transform dCTP into dATP in the protein environment (for both wild type and R283A Pol β mutant) and in solution. Nucleobase atoms (pyrimidine of dCTP and purine of dATP) were treated with soft-core potentials,[104] while interactions of sugar and phosphate atoms were interpolated linearly. Each transformation was carried out in 12 windows with lambda values of 0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, 0.99078. Initial configurations for each window were obtained from a preliminary NPT 1 ns run at λ=0.5, which allows to establish H-bonded base pairs for both dCTP and dATP, followed by a further 1 ns equilibration at the corresponding lambda value. We accumulated 10 ns NPT simulation for each window. The free energy change associated with each transformation

was estimated via thermodynamic integration. The estimates of the ΔG for the transformation in the protein environment and in solution were combined to provide an estimate of the relative binding affinity of dATP versus dCTP (ΔΔG). Each transformation was repeated 10 times, allowing estimating the associated errors. In addition, a Hamiltonian replica exchange scheme was

used to improve convergence during the alchemical free-energy calculations.[131]

Similar approaches have been previously used to study tautomeric G-T mismatches in Pol lambda,[132] Protein-DNA complexes,[133] Waston-Crick and Hoosteng base paring.[134]

4. Evaluating the performance of Alchemical Transfer Method (ATM) in ribonucleotide discrimination

This project was done in collaboration with Gianni De Fabritiis from the Univeristy Pompeu Fabra in Barcelona

4.1. Introduction.

Alchemical Free Energy (AFE) calculations have emerged as a highly effective tool for predicting binding affinities in drug discovery, particularly during hit-to-lead and lead optimization stages. These methods leverage non-physical, "alchemical" pathways to compute free energy differences between related molecules using approaches such as Free Energy Perturbation (FEP), Thermodynamic Integration (TI), and the Alchemical Transfer Method (ATM).[103,106,107] The implementation of these methods in various software packages including AMBER[135], GROMACS,[136] OpenMM,[137] and Schrödinger FEP+[138,139], typically achieve accuracies of ~1.0 kcal/mol RMSE across diverse protein-ligand systems.

For example, the Wang et al.[140] dataset encompasses over 330 ligand perturbations across eight diverse protein systems (BACE, CDK2, JNK1, MCL1, p38, PTP1B, thrombin, and TYK2), representing different binding sites, druggability, and chemical space coverage. Community-wide challenges like the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL),[141] particularly its host-guest benchmark sets, and the Drug Design Data Resource (D3R) Grand Challenges have further demonstrated AFE reliability across challenging systems including G-protein coupled receptors (GPCRs), ion channels, and membrane proteins.[142] Recent applications have successfully extended to fragment-based drug design, covalent inhibitors, and

even challenging targets like protein-protein interactions, establishing AFE as a useful computational tool for pharmaceutical research.

Despite this success, the application of AFE to nucleic acid systems presents persistent and significant challenges, primarily due to "accuracy gaps" in current force fields. The inherent flexibility of the sugar-phosphate backbone, with its six torsional angles per nucleotide, makes both accurate parameterization and effective conformational sampling extremely difficult. The polyanionic nature of DNA and RNA makes their dynamics highly sensitive to the surrounding solvent and ionic environment. This decreases the accuracy of the water model and the treatment of ions, especially divalent cations like Mg^{2+}, where electronic polarization effects become critical[94]. When combined with the slow conformational transitions common in nucleic acids, these force field inaccuracies can lead to poor convergence and high statistical errors.[105]

The discrimination between ribonucleotides (rNTPs) and deoxyribonucleotides (dNTPs) by DNA polymerases represents both a critical biological process for maintaining genomic integrity and an ideal test case for evaluating the capabilities and limitations of alchemical free energy calculations in nucleic acid systems. This selectivity, which is essential for proper DNA replication and cellular function, is primarily governed by the presence or absence of a single hydroxyl group at the 2′ position of the ribose sugar, a minor structural difference that produces profound changes in binding affinity and incorporation efficiency (Figure 17A). This selectivity mechanism is exemplified by well-characterized systems such as DNA Polymerase β (Pol β) and the bacteriophage RB69 polymerase, which have evolved highly effective molecular strategies to discriminate against ribonucleotide (rNTP) incorporation during DNA synthesis (Figure 17 B-C).

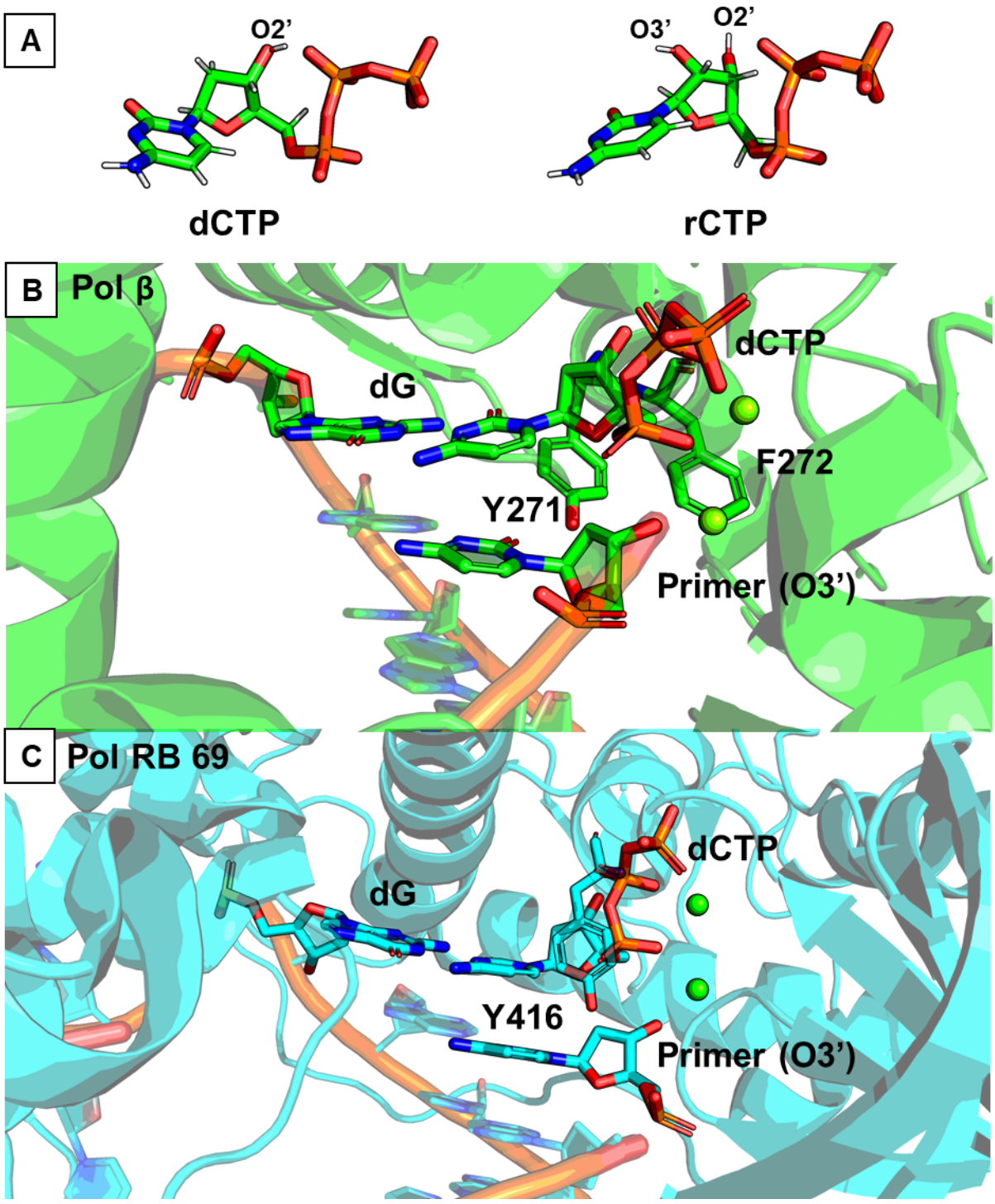


Figure 18 (A) Chemical structures of dCTP and rCTP showing the critical difference at the 2′ position: dCTP lacks the hydroxyl group (O2′) while rCTP contains it, creating the structural basis for polymerase discrimination. The 3′-OH position that participates in nucleophilic attack during catalysis is also indicated. Active site structures of (B) DNA polymerase β (green) and

(C) RB69 DNA polymerase (cyan) in ternary complexes with template dG, incoming dCTP substrate, and Mg2+ ions (green spheres). Both polymerases position critical tyrosine residues (Y271 in Pol β, Y416 in RB69) adjacent to the 2′ position of the incoming nucleotide, where steric clashes with the ribose 2′-OH group prevent rNTP incorporation. The primer 3′-OH position and neighboring F272 residue in Pol β is highlighted.

In Pol β, an X-family polymerase, the discrimination mechanism centers on residue Y271, where the backbone carbonyl oxygen creates an unfavorable steric interaction with the incoming rNTP's 2′-hydroxyl group (Figure 1B). This steric clash effectively destabilizes rNTP binding relative to the cognate dNTP substrate. Importantly, the Y271 side chain simultaneously forms a critical hydrogen bond to the primer terminus, helping to properly position the substrate for catalysis.[38] The specificity of this mechanism has been confirmed through systematic mutagenesis studies, which demonstrate that the Y271A mutant loses discrimination capability while the neighboring F272A mutant maintains normal selectivity, indicating the precise structural requirements for effective discrimination. The B-family RB69 polymerase uses the aromatic side chain of Y416 to create a direct steric block against the 2′-OH group of incoming rNTPs (Figure 1C). Mutagenesis of this residue to alanine (Y416A) dramatically compromises selectivity, allowing rNTPs to be incorporated with much higher efficiency than in the wild-type enzyme.[143] Previous computational studies by Yoon and Warshel successfully elucidated the mechanistic basis of ribonucleotide discrimination using EVB/FEP calculations, demonstrating that steric interactions govern selectivity in DNA polymerases while electrostatic effects dominate in RNA polymerases.[40]

Here, we evaluate the performance of the Alchemical Transfer Method (ATM) for predicting relative binding free energies in these well-characterized nucleotide discrimination systems. We investigate the impact of force field choice by comparing general small-molecule and DNA/RNA force fields during binding affinity changes related to polymerase mutations. We then explore

challenging nucleotide analogs such as dedoxynucleoside triphosphates (ddCTP) and therapeutic compounds like arabinonucleoside triphosphates (araCTP). This work identify and characterize the current advantages and limitations of force field representations in nucleic acid systems. Then, we propose how emerging Neural Network Potential/Molecular Mechanics (NNP/MM) methodologies can provide a robust solution to overcome the accuracy limitations of current classical force fields in AFE calculations for nucleic acid systems.

4.2. Results

4.2.1. SHAW Force fields outperform general purpose force field GAFF2 predicting the impact of single point mutations in ribonucleotide discrimination.

We initially evaluated the capability of the GAFF2 general force field to reproduce experimental ribonucleotide discrimination patterns in both DNA Polymerase β and RB69 polymerase systems. Alchemical transformations from dCTP to rCTP were performed across all mutant systems: wild-type, Y271A, Y271F, and F272A for Pol β, and wild-type, Y416A, and Y416F for RB69 polymerase. All mutants, are in direct contact with the incoming nucleotide (XCTP) and the O3' from the primer DNA strand that perform the nucleophilic attract during Polymerase catalysis.

The GAFF2 force field yielded an overall mean absolute error (MAE) of 1.51 kcal/mol when compared to experimental relative binding free energies derived from kinetic data (Figure 18, top panel). While several systems showed reasonable agreement with experimental values. Specifically the wild-type, Y271F, and F272A systems for Pol β and the wild-type and Y416F systems for RB69 (combined MAE of 0.93 kcal/mol). GAFF2 systematically failed to reproduce

the discrimination behavior of the most critical mutants. The Y271A mutant in Pol β and the Y416A mutant in RB69, both known experimentally to lose ribonucleotide discrimination capability, showed large deviations from experimental values with MAEs of 3.56 kcal/mol and 2.32 kcal/mol, respectively. The large deviation erroneously suggest a preference for rCTP over dCTP in DNA polymerases, showing significant limitations in GAFF2's ability to capture the subtle energetic differences that govern nucleotide selectivity.

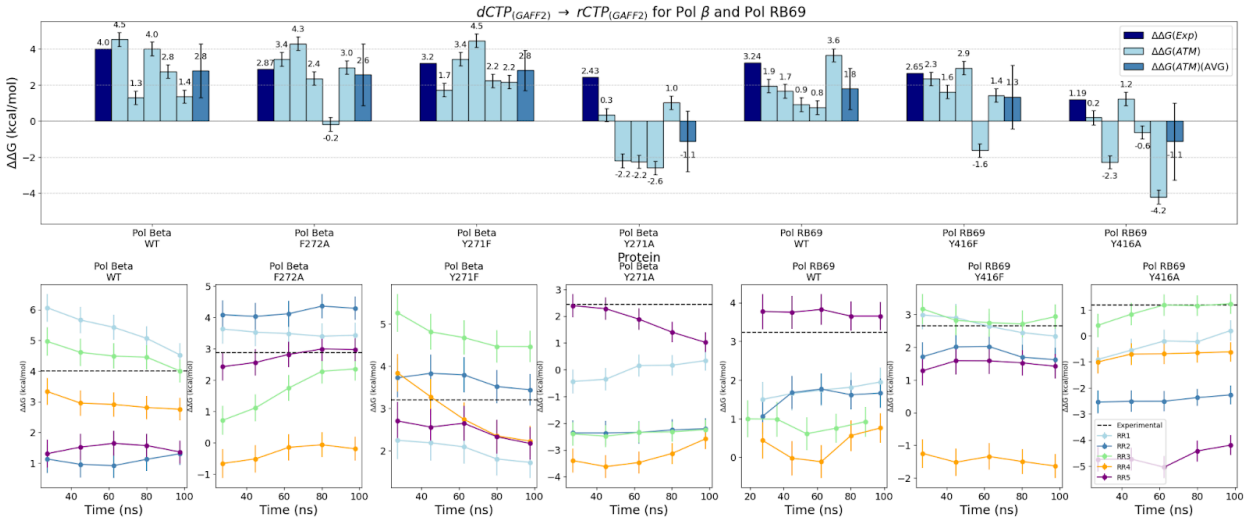


Figure 19 Performance of the GAFF2 force field in predicting ribonucleotide discrimination. (Top) Comparison of experimental ΔΔG values (dark blue) with ATM-calculated averages (dark teal). GAFF2 systematically fails to predict the loss of discrimination in the Y271A and Y416A mutants. (Bottom) Convergence plots for the five individual replicas, showing high variance and poor convergence for several systems, particularly the key mutants.

To address the limitations observed with GAFF2, we performed all calculations using the SHAW force field, which has been specifically parameterized for nucleotide systems. This specialized force field demonstrated substantial improvements across all metrics evaluated (Figure 19). These simulations showed an overall improvement in the mean MAE (0.71 kcal/mol versus 1.51 kcal/mol from GAFF2 calculations). The Y271A and Y416A mutants in Pol β and Pol RB69 showed lower MAE values (1.3 and 0.93 kcal/mol, respectively). Importantly, SHAW force field correctly predicted a preference for dCTP over rCTP in both DNA polymerases. The calculated

ΔΔG values were 1.13 ± 1.69 kcal/mol for Pol β and 0.26 ± 1.43 kcal/mol for Pol RB69, compared to experimental values of 2.43 and 1.19 kcal/mol, respectively.

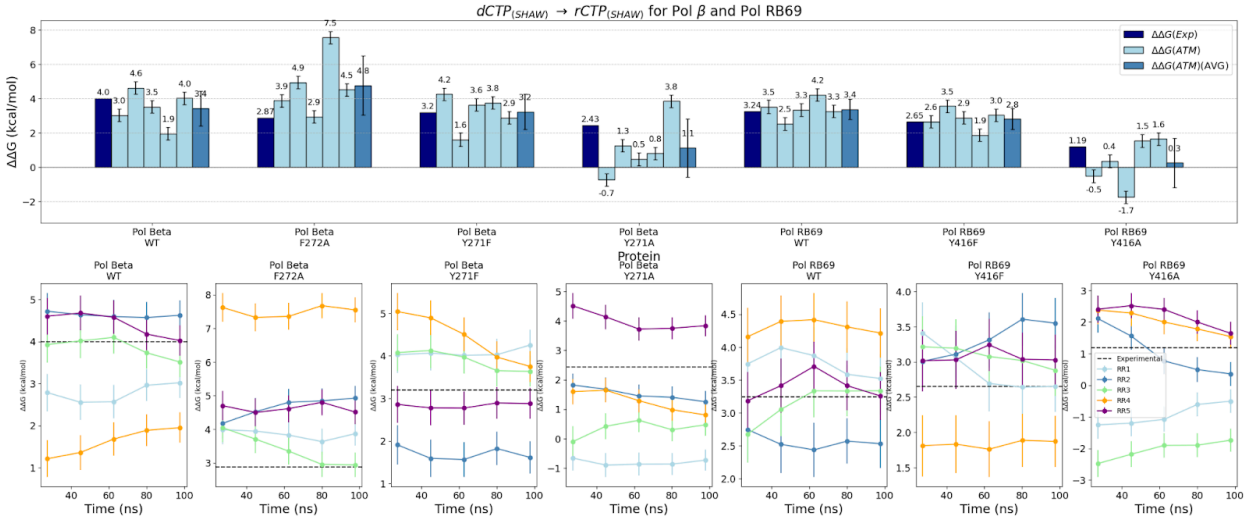


Figure 20. Performance of the SHAW force field in predicting ribonucleotide discrimination. (Top) Comparison of experimental ΔΔG values (dark blue) with ATM-calculated averages (dark teal). SHAW FF predicts the loss of discrimination in the Y271A and Y416A mutants with better accuracy. (Bottom) Convergence plots for the five individual replicas, showing high variance and poor convergence for several systems, particularly the key mutants.

To ensure that our calculated binding affinities reflected realistic protein-nucleotide interactions, we performed detailed analysis of active site geometry throughout the simulations (Table S 5 and Figure S 18 and Figure S 19). Key catalytic distances such as the nucleophilic O3′-Pα distance, metal coordination distances (O3′-MgA and Pα-MgA), and base pair hydrogen bonding remained stable and consistent with crystallographic observations.

4.2.2.SHAW force field do not correctly predict relative binding affinities in nucleotide analogs.

After evaluating how the ATM method could predict the relative effects of mutations in residues that directly contact the incoming nucleotide, we shifted our focus to investigating how subtle

structural modifications within the sugar ring affect binding affinity. Specifically, we examined how variations in the hydroxyl groups present on the ribose sugar influence polymerase selectivity and binding energetics.

In addition to the dCTP to rCTP transformation, we extended our analysis to include two important nucleotide analogs: dideoxynucleotide cytidine triphosphate (ddCTP) and cytarabine triphosphate (araCTP). These two nucleotide variants differ only in their sugar moiety. ddCTP lacks both 2′ and 3′ hydroxyl groups, and araCTP features hydroxyl groups at both the 2′ and 3′ positions but with an inverted stereochemical configuration at the 3′ carbon in comparison with rCTP. These nucleotide analogs serve distinct roles in both research and therapeutic applications. ddCTP is widely employed in biochemical analyses of polymerase reaction mechanisms, as its lack of a 3′-OH group results in chain termination upon incorporation, making it a valuable tool for studying DNA synthesis pathways and polymerase kinetics. In contrast, araCTP functions as a competitive inhibitor of DNA polymerases.[144,145]

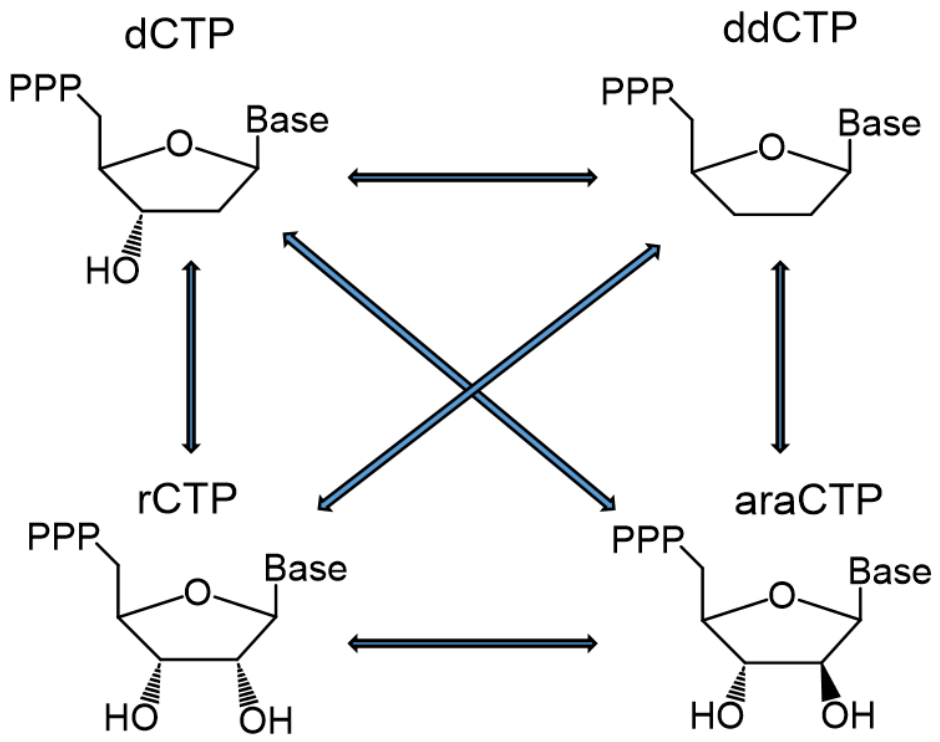


Figure 21. Thermodynamic cycle for alchemical transformations of nucleotide sugar modifications. Schematic representation of the four nucleotide substrates studied: dCTP (deoxyribose with 3′-OH only), rCTP (ribose with both 2′-OH and 3′-OH), ddCTP (dideoxyribose lacking both hydroxyl groups), and araCTP (arabinose with both hydroxyl groups but inverted 3′ stereochemistry). Arrows indicate possible alchemical transformation pathways that can be used to calculate relative binding free energies between any pair of nucleotides. PPP represents the triphosphate group, and "Base" indicates the cytosine nucleobase, which remains constant across all transformations to isolate the energetic contributions of sugar modifications.

As shown in Figure 21, transformations involving ribonucleotides (dCTP ↔ rCTP and dCTP ↔ araCTP) showed excellent agreement with experimental values, achieving an average MAE of 0.59 kcal/mol with low variability across replicas (standard deviations of 0.6-1.1 kcal/mol). These results demonstrate that when specialized force fields are properly applied, alchemical methods can accurately predict the energetic consequences of hydroxyl group modifications in nucleotide systems. In contrast, simulations containing dideoxynucleotides (ddCTP) showed

large errors and variability (average MAE of 4.83 kcal/mol). This reflects fundamental limitations in classical force field representations. When subtle changes in the molecular structure are made, there is a disruption in the carefully torsional and electrostatic parameters that govern sugar pucker and local conformational preferences. The extensive effort invested in specialized force fields is not generalizable to these changes.

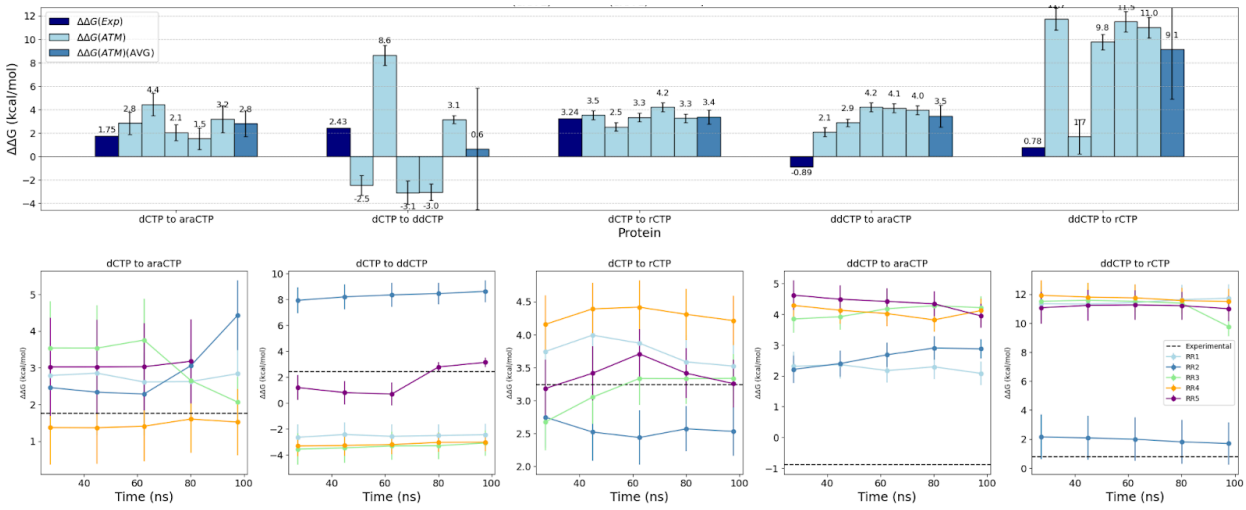


Figure 22. ΔG estimation for the transformation of dCTP into rCTP,ddCTP and araCTP for all replicas and average value using SHAW as FF for the incoming nucleotide. Convergence plot of all the replicas.

Both Pol RB69-rCTP and Pol RB69-araCTP complexes proved structurally unstable across all simulation conditions tested. Despite extensive equilibration protocols and extended unbiased molecular dynamics simulations (>100 ns), these systems consistently developed severe steric clashes between the additional hydroxyl groups and active site residues. Detailed structural analysis revealed that the extra hydroxyl group in arabinose creates steric conflicts with the Y416 aromatic ring and the N564 side chain (Figure 22).

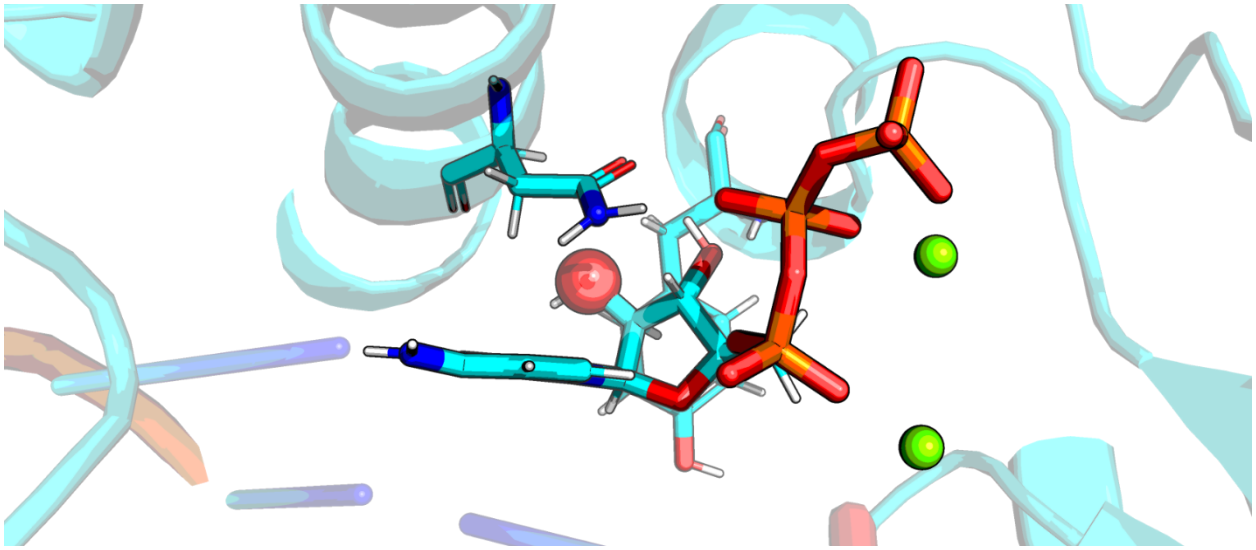


Figure 23. Structural representation showing the severe steric conflicts that arise when araCTP (with arabinose sugar) is positioned in the RB69 active site. The incoming nucleotide is displayed in stick representation with the cytosine base (cyan), arabinose sugar (cyan carbons), and triphosphate groups (red and orange). The additional 3′-hydroxyl group in the S configuration (highlighted with pink sphere) creates steric clashes with the aromatic ring of Y416 and the side chain of N564 (shown as cyan sticks). Mg^{2+} ions are represented as green spheres, and the protein environment is shown as cyan ribbons with DNA strands as blue and orange ribbons.

4.3. Future work.

The substantial errors encountered for dideoxynucleotide transformations and the absence of specialized force fields for many nucleotide variants highlight fundamental limitations in current classical molecular mechanics approaches. Neural Network Potentials (NNPs) offer a plausible solution to address the modeling of these molecules. These machine learning-based approaches can achieve quantum mechanical-level accuracy while maintaining computational efficiency suitable for free energy calculations. The hybrid NNP/MM methodology, which treats the ligand with an NNP while maintaining the protein and solvent environment with classical mechanics,

can be directly applied to the ATM methodology and have been successfully used in RBFE contexts.[43,108,110]

A critical requirement for successful application of NNPs to nucleotide systems is the development of comprehensive, high-quality training datasets. Existing general-purpose NNPs often show dramatically reduced accuracy when applied to molecular systems that are poorly represented in their training data, with charged molecules presenting particular challenges. Given that nucleotides are not commonly encountered in typical protein-ligand binding contexts, we have initiated the creation of a specialized dataset focused specifically on nucleotide chemistry. Our training dataset encompasses diverse structural variations including different phosphorylation states (mono-, di-, and triphosphates), sugar modifications (ribose, deoxyribose, arabinose, and other therapeutically relevant variants), and modified nucleobases (Figure 23). This comprehensive approach ensures adequate representation of the conformational and chemical space relevant to nucleotide-protein interactions while properly accounting for the high charge density characteristic of nucleotide triphosphates.

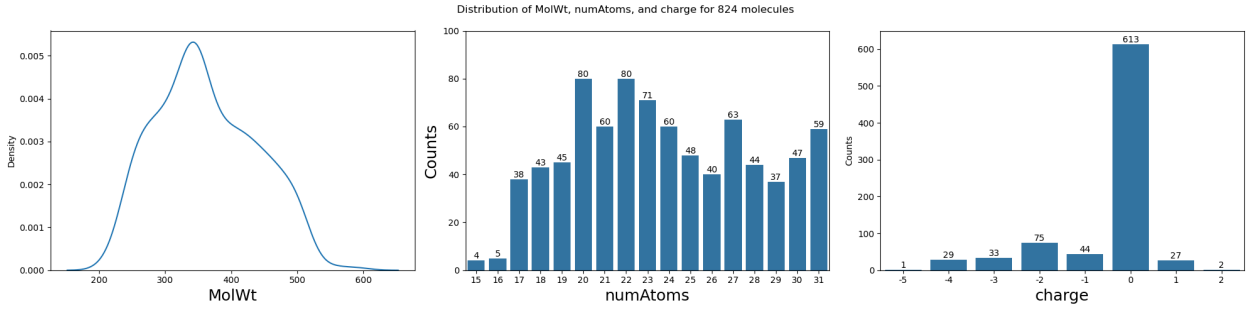


Figure 24. Molecular property distributions in the nucleotide training dataset. Statistical analysis of 824 nucleotide and nucleotide-analog molecules included in the specialized training dataset for neural network potential development. Molecular weight distribution showing coverage from approximately 200 to 600 Da. Distribution of total atom count per molecule, demonstrating systematic coverage of small to medium-sized nucleotides with the majority containing 20-25 heavy atoms. Charge distribution revealing predominantly neutral molecules alongside charged species ranging from -4 to +2, reflecting the diverse phosphorylation states and chemical

modifications included in the dataset. Some example of the molecules included in this training set are shown in Figure S 21. Example of nucleotide modifications included in the NNP dataset.Figure S 21.

Once these improved versions of NNP/MM for nucleotides are trained, calculations will have an increase in computational time but are expected to show a significant increase in accuracy. The additional computational cost is justified by the need for reliable predictions in these challenging systems where classical force fields have shown persistent limitations.

4.4. Conclusions

This study assessed the accuracy of Alchemical Free Energy (AFE) calculations in predicting the subtle binding affinity differences that govern nucleotide selection in DNA polymerases. This process is critical for maintaining genomic integrity and represents a challenging test for computational methods. Our findings demonstrate that the GAFF2 force field fails to accurately model the impact of single point mutations on ribonucleotide discrimination, particularly for key mutants like Pol β Y271A and Pol RB69 Y416A, which are critical for selectivity. The specialized nucleotide force field SHAW showed overall improvement and correctly predicted the preference for dCTP over rCTP in these mutants. However, significant inaccuracies persisted, especially in transformations involving dideoxynucleotides (ddCTP), which showed large errors (MAE of 4.83 kcal/mol) and high variability. These persistent errors underscore the limitations of classical force fields in describing the nuanced conformational and electrostatic properties of nucleotide analogs. The extensive parameterization efforts invested in specialized force fields are not generalizable to structural variations such as the absence of hydroxyl groups in dideoxynucleotides.

To address these challenges, future work will implement a hybrid Neural Network Potential/Molecular Mechanics (NNP/MM) methodology. This approach promises quantum mechanical-level accuracy for complex systems. We will develop a specialized training dataset of nucleotide-like molecules to ensure the NNP is well-suited for these chemically diverse and highly charged systems. While this approach will increase computational cost, it aims to achieve more accurate and reliable predictions of nucleotide binding and polymerase selectivity, which current classical force fields cannot provide.

4.5. Computational details.

We start from matched dG:dCTP ternary complexes PDB 4LVS (1.85 A) and PDB 3NCI (1.79A) from Pol β and Pol RB69, respectively. Pdb2pqr software was used to assign protonation states corresponding to pH 7.4.[125] Solvation was performed in cubic boxes of 12 A and Na+ and Cl- ions were added to simulate a 150 uM concentration. W3DNA server[126] was used to match the sequence used in kinetic experiments (Template (3′-CGACTACGCCGCAGCC- 5′ and 3′-GGCTGGTCGGAACGTTTTTT-5′) for both Pol β and PolRB69). Mutations for both polymerases were introduced using PyMol.[127]

The protein and DNA were described using AMBERFF14SB[86] and parmbsc1[128] force fields, respectively. For Mg2+ ions, van der Waals parameters by Allner et al. were used.[129] Joung−Cheatham parameters for K+ and Cl− ions were used.[146] The TIP3P water molecule was adopted. Charges for dCTP,rCTP,ddCTP and araCTP were taken or calculated from the REDDB database.[130] Topologies were prepared with Ambertools.[147]

Equilibration protocal was based on previous validations of the ATM methodology.[43] Initially, the system's potential energy is minimized using a standard convergence-based algorithm. This is followed by a multi-stage equilibration process conducted with a 1 fs time step. The first stage is a thermalization phase in the NVT ensemble, where the system is gradually heated from an initial temperature of 50 K to the final target temperature of 300 K. This heating occurs over 150,000 steps (150 ps), divided into 30 cycles of 5,000 steps each, with the temperature increased linearly after each cycle. Temperature is maintained using a Langevin integrator. Next, an NPT equilibration is performed for another 150,000 steps (150 ps) at 300 K and 1 bar, during which a Monte Carlo barostat is applied with a frequency of 25 steps to allow the system volume to relax. Following this, the system undergoes a final NVT equilibration for an additional 150,000 steps (150 ps). After this initial structural relaxation, an alchemical preparation is carried out. An annealing simulation of 250,000 steps (250 ps) is run in the NVT ensemble, during which the alchemical coupling parameter λ is linearly increased from 0.0 to 0.5. Finally, to ensure the system is well-equilibrated at the intermediate alchemical state, a concluding NVT equilibration is run for 150,000 steps (150 ps) with λ held constant at 0.5. This entire procedure, totaling 850 ps (0.85 ns), generates the final starting configuration for subsequent production simulations.

Parallel replica exchange molecular dynamics was employed along the λ-coordinate using the ATMForce potential implementation in OpenMM 8.2.[106,108] The λ-schedule and softcore parameters were optimized based on previous studies to ensure smooth free energy profiles and adequate phase space overlap between adjacent windows (parameters detailed in Supplementary Table S 4. lambda scheduling and softcore parameters used in all transformations). Five independent replicas were executed for each transformation to ensure statistical reliability, with

each replica consisting of 22 λ-windows spanning the complete alchemical pathway. Production simulations used a 2 fs timestep and were run for 12 ns per window (264 ns total per replica). To enhance convergence and maintain proper ligand alignment during alchemical transformations, a harmonic restraining potential was applied between corresponding atoms of the two nucleotide substrates. This restraint was designed to preserve the relative position and orientation of the nucleotides while allowing necessary conformational flexibility for the sugar moiety transformations.

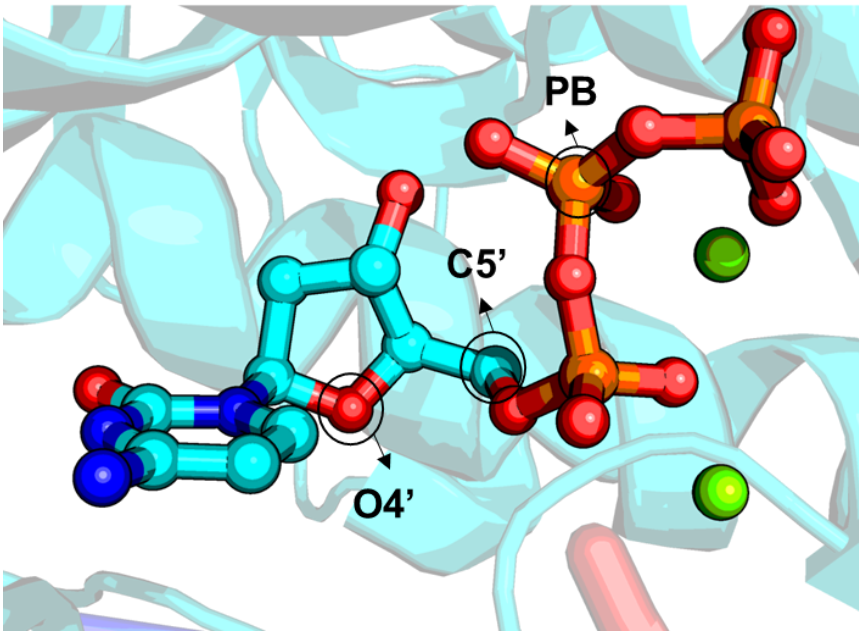


Figure 25. Structure of dCTP in Pol RB69. In circles denoted the atoms used for the positional restrain applied to maintain orientation during transformation. Molecules of rCTP,ddCTP and araCTP also had the restrain in the same atoms, aiming to cover relative position of the sugar and the phosphate. Base pair interaction of the nucleobase with the template dG were monitored in all simulations. Hydrogens not shown for clarity.

5. A Classification-Based Convolutional Neural Network for Fragment Prediction in Structure-Based Drug Design

This project was done in collaboration with Luigi Zanovello and Massimiliano Pontil from IIT.

5.1. Introduction

Structure-Based Drug Design (SBDD) has undergone a significant transformation with the integration of machine learning approaches, particularly deep learning methods that can leverage the wealth of three-dimensional structural information available from X-ray crystallography, NMR spectroscopy, and computational structure prediction tools.[72,148,149] These computational advances allow exploring solutions beyond traditional physics-based approaches and empirical scoring functions, focusing on data-driven methods that can learn complex patterns from large datasets of protein-ligand interactions.[150–152] The application of convolutional neural networks to voxelized protein binding sites,[74,153,154] graph neural networks to molecular representations,[155,156] and transformer architectures to sequence-structure relationships has opened new possibilities for automated molecular design and optimization.[157,158] Three key architectural innovations have opened new possibilities for automated molecular design and optimization. Convolutional neural networks applied to voxelized protein binding sites treat binding environments as 3D grids where spatial relationships between atoms are preserved through convolution operations. Graph neural networks process molecular representations by encoding atoms as nodes and bonds as edges, capturing connectivity patterns and chemical relationships. Transformer architectures applied to sequence-structure relationships leverage attention mechanisms to identify long-range dependencies between amino acid sequences and their corresponding structural motifs.[113,159,160]

While generative AI models for molecular design have shown promise, they face significant challenges that limit their practical deployment. These include the generation of chemically

invalid structures such as molecules with impossible valences, violations of chemical stability rules, difficult interpretability of failure cases, and unclear evaluation metrics that make it challenging to assess model reliability. Models such as variational autoencoders, generative adversarial networks, and autoregressive transformers have demonstrated these limitations in practice.[161–163] Classification approaches address these fundamental limitations by ensuring chemical validity through selection from known fragment libraries, providing clearer interpretability of predictions through explicit fragment identification, and enabling straightforward evaluation through standard classification metrics including accuracy, F1-score, area under the curve, precision, recall, and top-k accuracy measures. This framework is particularly crucial for practical deployment, where medicinal chemists require both accurate and interpretable suggestions for synthesis planning.

Tools like DeepFrag, based on the classification approach, offer distinct advantages in terms of interpretability and chemical validity, as they operate within the constrained space of known molecular entities rather than attempting to generate entirely novel structures de novo.[76,77] Despite the advantage of this architecture, such as a well-defined loss function, current methodologies face several limitations, including reliance on indirect prediction schemes through molecular fingerprints, single fragmentation strategies that may miss important chemical diversity, and insufficient consideration of protein structural diversity and druggable pocket characteristics. Without addressing the inherent biases in the PDB, where certain protein families are overrepresented, models risk memorizing family-specific features rather than learning generalizable patterns.

To address these fundamental challenges, we present a comprehensive classification-based methodology for fragment-based drug design, where we predict the most probable chemical fragment given a voxelized protein environment. This method integrates three critical innovations. Our approach combines direct fragment class prediction that eliminates the need for indirect fingerprint-based schemes, dual fragmentation algorithms that merge cyclic and aromatic scaffolds with functional group decomposition to ensure comprehensive chemical space coverage by capturing both rigid ring systems and flexible heteroatom-containing moieties that single-method approaches miss, and systematic protein clustering with physicochemical pocket filtering to eliminate family-specific bias while ensuring fragments occupy well-defined binding

cavities through MMseqs2-based sequence clustering at 50% identity combined with solvent-accessible surface area filtering. (Figure 25)

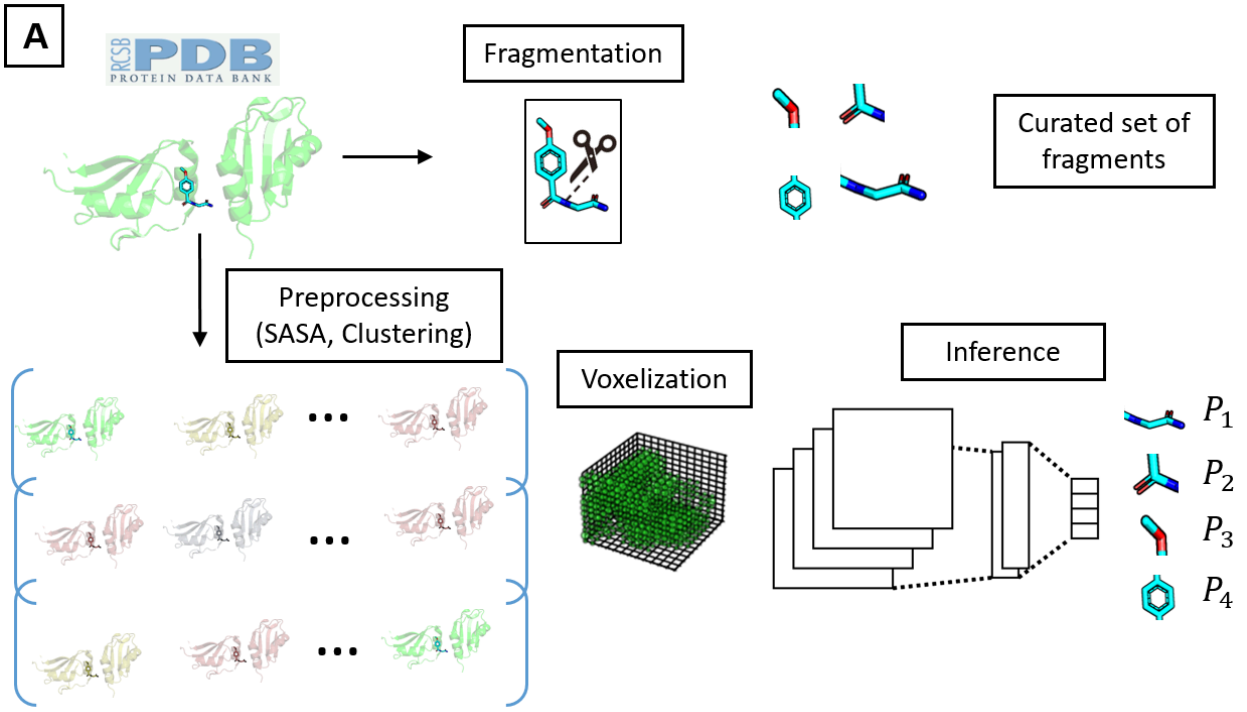


Figure 26. Protein-ligand complexes from the PDB undergo dual fragmentation (RDKit scaffold network and Ertl functional group analysis) to generate a curated fragment library. The protein structures are preprocessed through SASA filtering (threshold <0.25) to retain fragments in buried binding pockets and sequence clustering (50% identity, MMseqs2) to reduce bias from overrepresented protein families. Filtered complexes are voxelized into 12×12×12 Å³ grids encoding the local atomic environment, which serve as input for a 3D CNN that predicts appropriate fragment classes (P_1, P_2, P_3, P_4) for the given binding site.

Through systematic optimization of these integrated approaches, our methodology demonstrates robust generalization capabilities that address the critical challenge of transferability to novel protein families. Our final model achieves 67% top-1 accuracy and 80% top-5 accuracy across 30 chemically diverse fragment classes, representing performance levels competitive with existing fragment-based approaches while providing superior interpretability, with minimal performance degradation when evaluated on completely novel protein families through sequence-based splitting. Validation on clinically relevant targets including SARS-CoV-2 3CL

protease inhibitors, ERK2 kinase inhibitors, and HCV NS5B polymerase inhibitors reveals both the method's strengths in recognizing rigid aromatic scaffolds and core heterocyclic systems, and critical limitations in identifying flexible aliphatic regions, complex functional groups like sulfonamides, and conformationally diverse linker systems. These findings establish a clear roadmap for future development while demonstrating the current method's utility for specific fragment classes within structure-based drug design applications.

5.2. Results

5.2.1. Dataset and fragment library statistics

During the PDB filtering process, we systematically processed 215538 total PDB entries to identify high-quality protein-ligand complexes suitable for our fragment-based classification approach, ultimately collecting 105,131 PDB structure complexes that met our stringent resolution criterion of better than 2.5 Å, which yielded a total of 30,863 unique ligands based on their canonical SMILES representations after removing redundant molecules across different crystal structures. We considered as ligands every small molecule entry present in each PDB structure, a definition that initially encompassed not only the intended drug-like compounds and natural product inhibitors but also cocrystallized solvent molecules. Then, we applied additional chemical validity filters to remove ions such as Na+, K+, Cl-, $SO4^{2-}$, and $PO4^{3-}$ that lack the covalent bonding patterns necessary for fragment-based analysis. Additionally, we filter out SMILES molecules retrieved from the RCSB API[164] that failed proper chemical processing with RDKit due to issues such as invalid valences, ambiguous stereochemistry assignments, or incomplete atomic coordinates, ensuring that our dataset contained only chemically well-defined molecules.

The molecular properties from the final dataset showed molecular weight ranging from approximately 100 to 2,500 Da with a mean of 372.79 ± 175.26 Da, the synthetic accessibility (SAS) scores calculated using the Ertl and Schuffenhauer method[165] ranging from 1 to 8 with a mean of 3.17 ± 1.09, along with additional physicochemical descriptors such as lipophilicity (MolLogP: 2.19 ± 2.55), hydrogen bond donors (2.58 ± 2.18) and acceptors (5.45 ± 3.64), rotatable bond counts (5.66 ± 5.26), nitrogen and oxygen atom counts (6.88 ± 4.44), NH/OH

group counts (2.93 ± 2.51), and ring count (2.86 ± 1.76 rings per molecule), are visualized in Figure 26, revealing that our filtered dataset captures a diverse chemical space representative of drug-like molecules

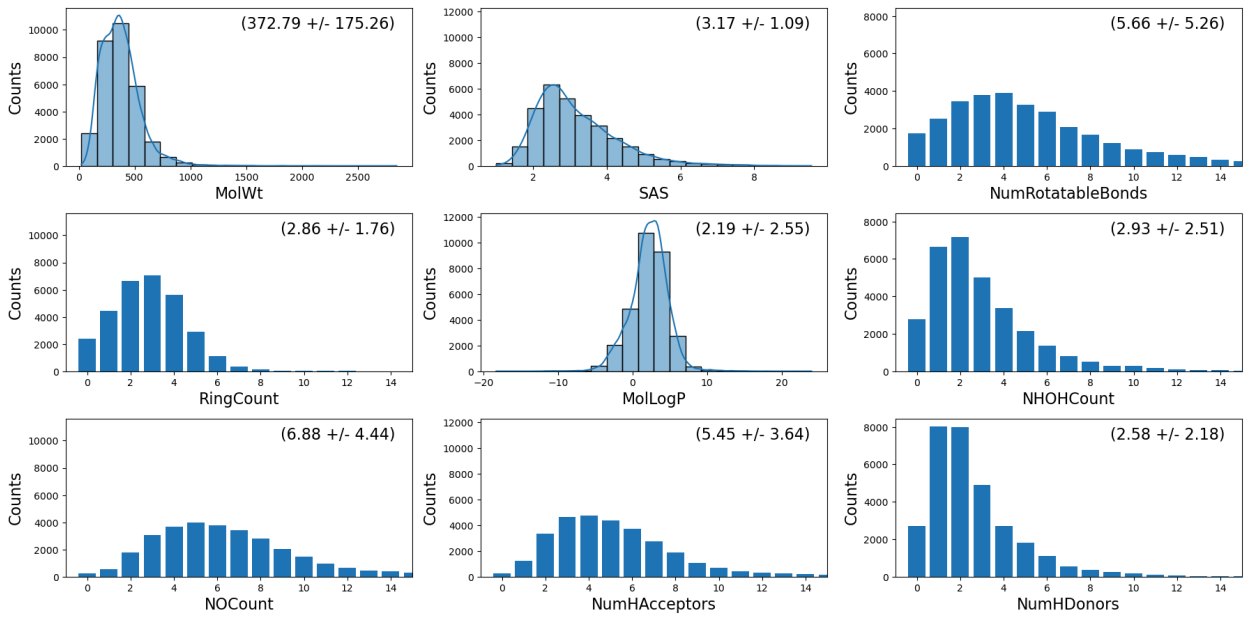


Figure 27. Distribution of molecular properties for the filtered PDB ligand dataset. Histograms showing the distribution of key physicochemical properties across 30,863 unique ligands after initial filtering steps. Properties include molecular weight (MolWt), synthetic accessibility score (SAS), number of rotatable bonds (NumRotatableBonds), ring count (RingCount), molecular lipophilicity (MolLogP), counts of nitrogen and oxygen atoms (NOCount), hydrogen bond acceptors (NumHAcceptors), hydrogen bond donors (NumHDonors), and NH/OH groups (NHOHCount). Mean values and standard deviations are indicated in parentheses for each property.

Following this preliminary characterization, we fragmented all 30,863 unique ligands using our dual fragmentation strategy to obtain complementary sets of molecular substructures: functional groups enriched in heteroatoms through the Ertl algorithm, which systematically cleaves bonds between carbons and heteroatoms to isolate chemically reactive moieties, and cyclic and aromatic systems through the Scaffold Network implementation in RDKit,[166] which preserves ring systems and their connecting linkers as coherent structural units. The distribution of the resulting fragments across our dataset reveals striking frequency variations, with the most common Ertl fragments including simple hydroxyl groups (CO: 80,835 in all PDBs), methoxy ethers (COC: 33,798 PDBs), and carboxylic acids (CC(=O)O: 23,451 PDBs), while the Murcko

decomposition identified prevalent scaffolds such as benzene (C1=CC=CC=C1: 32,812 PDBs), pyrimidine (C1=CN=CN=C1: 27,511PDBs), and imidazole (C1=CN=CN1: 21,189) ring systems, demonstrating both the chemical diversity captured and the inherent imbalance that necessitates careful dataset curation due to the presence of fragments with no so many examples in our dataset (Figures S1-S2).

5.2.2. Impact of Protein Clustering and SASA Filtering on Dataset Size and Diversity

With the fragments selected from both functional groups and cyclic-aromatic groups, we performed systematic filtering to create a well-curated dataset suitable for machine learning, implementing both solvent accessibility filtering and protein sequence clustering to address inherent biases in the PDB. The SASA-based filtering was applied first, using a threshold of 0.25 for Ertl fragments and 0.5 for RDKit scaffolds networks to select fragments occupying well-defined binding pockets. The higher relative value used in the for RDKit scaffolds networks aimed to capture more of the protein environment surroundings in drug-like molecules where hydrophobic cores allow hydrophilic fragments to interact but are not in direct contact with the protein aminoacids. The SASA filter change fragment distributions, with hydroxyl groups (CO) showing moderate retention from 98,259 to 80,835 structures (82.3%), while aromatic variants like cO retained only 7,063 from initial PDB entries, and methoxy groups (COC) reduced from 41,685 to 33,798 (81.1%), demonstrating that simple polar groups often interact with solvent-exposed regions rather than buried cavities. The ciclyc aliphatic/aromatic fragments showed generally higher retention rates, with benzene rings maintaining 32,812 from 34,267 structures (95.8%) and pyrimidine retaining 27,511 from 28,551 (96.4%), confirming that hydrophobic aromatic systems preferentially occupy buried protein pockets where specific molecular recognition occurs, while smaller heterocycles like cyclopropane (C1CC1) showed more modest retention from 1,549 to 1,422 structures (91.8%), as illustrated in Figure S 22-S23.

Following SASA filtering, we addressed the well-documented overrepresentation of certain protein families in the PDB through sequence-based clustering using MMseqs2[167] with a 50% identity threshold and 80% coverage requirement, which produced the most dramatic reduction in dataset redundancy. The hydroxyl fragment (CO) decreased from 80,835 SASA-filtered

structures to just 13,766 distinct protein clusters (83.0% reduction), methoxy groups fell from 33,798 to 7,294 clusters (78.4% reduction), and carboxylic acids from 23,451 to 5,519 clusters (76.5% reduction), effectively eliminating the bias where multiple crystal structures of the same or highly homologous proteins would otherwise dominate the training data. In the case of cyclic scaffolds. The clustering allow to reduce the big class imbalance by reducing high frequency fragments such as benzene and pymidine from ~32k and ~27k PDBs to 3193 and 4835 cluster, respectively.

5.2.3.Hyperparameters optimization

We produce the voxel for each protein-ligand with dimensions 12x12x12 Å and 3 channels, with a voxel size of 1 Å corresponding to the presence of carbon, oxygen, and nitrogen atoms for both the atoms of the protein receptor (RecOnly) and also the atoms of the molecule that are not part of the fragment (PatOnly, parent only). Specifically, the PatOnly input comprises the atoms from the parent ligand structure excluding the fragment itself, effectively representing the scaffold or immediate chemical context to which the fragment attaches. To address the rotational invariance inherent in 3D grid-based learning, we applied 24 discrete rotations to every input sample. The total dimension of the joint voxel is 12x12x12x6. For our initial optimization phase, we systematically explored the effects of voxel preparation methods, dataset balancing strategies, model architectures, and training parameters, using a focused setting of 10 representative fragment classes for rapid evaluation. We trained our three distinct CNN architectures (CNNBase_1C, CNNBase, and CNNScore) which, which operate by sliding 3D filters across the voxel grid to detect local spatial arrangements of atoms before pooling these features for classification. Among these, CNNBase_1C serves as a shallow baseline consisting of only a single 3D convolutional layer, whereas CNNBase utilizes three sequential layers to capture more complex representations using voxelized inputs with a binary representation (r=0) of three different channel combinations: receptor atoms only (RecOnly), atoms from the parent ligand excluding the fragment (ParentOnly), and a combination of both (Rec+Parent). Using a 12 Å grid size, both the Rec+Parent and ParentOnly configurations achieved high test accuracies, ranging from 0.84-0.85 for the simplest model to 0.91-0.93 for the more complex architectures. Conversely, the RecOnly models, which represent the only practical scenario for de novo prediction, yielded substantially lower performance, with test accuracies of only 0.40-0.42

(CNNBase_1C), 0.48-0.50 (CNNBase), and 0.53-0.54 for CNNScore (See Figure 27). This performance gap demonstrates that, without careful optimization, the models predominantly learned predictive features from the co-crystallized parent molecule's structure rather than the characteristics of the protein binding pocket. Recognizing this critical challenge, our subsequent efforts were exclusively focused on optimizing the RecOnly model to ensure it learns transferable binding site patterns, through a systematic exploration of data representation, dataset balancing, and model regularization.

5.2.3.1. Voxel Smoothing (r=1.75) Improves overall Accuracy

To better represent the continuous nature of atomic distributions and make the model more robust to minor positional variations, we investigated the effect of a smoothed voxel representation. The implementation of a Gaussian smoothing function with a radius of r=1.75 Å demonstrated consistent and significant accuracy improvements across all RecOnly models, confirming our hypothesis that this approach provides a more physically realistic input (For more details, see 5.5.9). The performance gains were: the test accuracy for CNNBase_1C increased from 0.40 to 0.51 (an absolute improvement of 0.11), CNNBase rose from 0.48 to 0.59 (also a 0.11 increase), and the more complex CNNScore improved from 0.53 to 0.59 (a +0.06 gain). This enhancement suggests that smoothing allows the convolutional filters to learn more generalizable features by mitigating the impact of slight conformational differences inherent in crystal structures. While this smoothing also conferred minor benefits to the ParentOnly and Rec+Pat models, these models still present an overfitting on the parent atom positions. (See Figure 27)

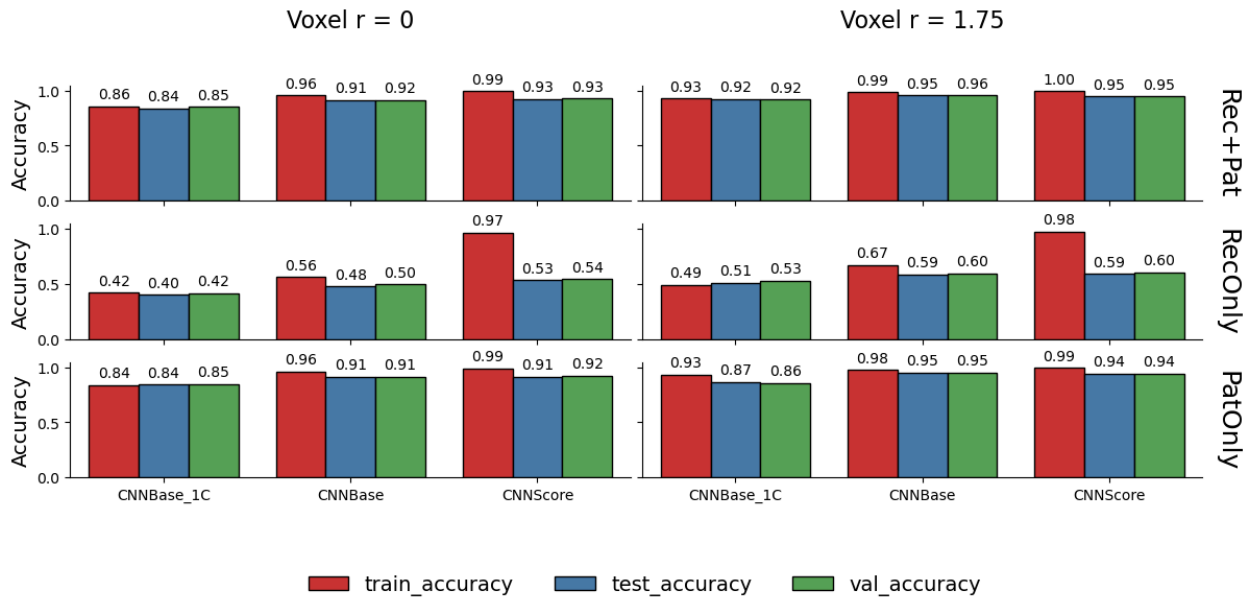


Figure 28. Impact of voxel smoothing on accuracy across different CNN Architectures and input configurations. The bar plots compare training (red), test (blue), and validation (green) accuracies for models trained with a binary voxel representation (left panel, r=0) versus a smoothed representation (right panel, r=1.75). Performance is evaluated for three input types: combined receptor and parent ligand (Rec+Pat), receptor only (RecOnly), and parent ligand only (PatOnly). The application of voxel smoothing (r=1.75) yields a consistent and significant improvement in test and validation accuracy for the RecOnly models, which is the most challenging and practically relevant scenario.

5.2.3.2 Effect of dataset balancing strategies on the overall accuracy.

Having established the optimal voxel representation through smoothing, we next addressed the inherent class imbalance in our fragment distribution, recognizing that even well-represented atomic environments cannot overcome severe data imbalance where some fragments appear in orders of magnitude more structures than others. Given this, we evaluated three distinct dataset preprocessing strategies. Unbalanced dataset**:** Using all available data after initial filtering, selecting one representative structure from each protein sequence cluster per fragment class. Balanced**:** A reduced dataset retaining only fragment classes appearing with minimum frequency thresholds, creating smaller but more balanced datasets. Unbalanced MAX10K**:** For each fragment class, protein-ligand complexes are first clustered by sequence similarity, then one structure is selected from each cluster. To address class imbalance where some fragments have many clusters while others have few, additional structures are sampled from already-used

clusters trying to reach 10,000 samples per class. This approach maximizes protein diversity by prioritizing different sequence clusters before resampling within clusters.

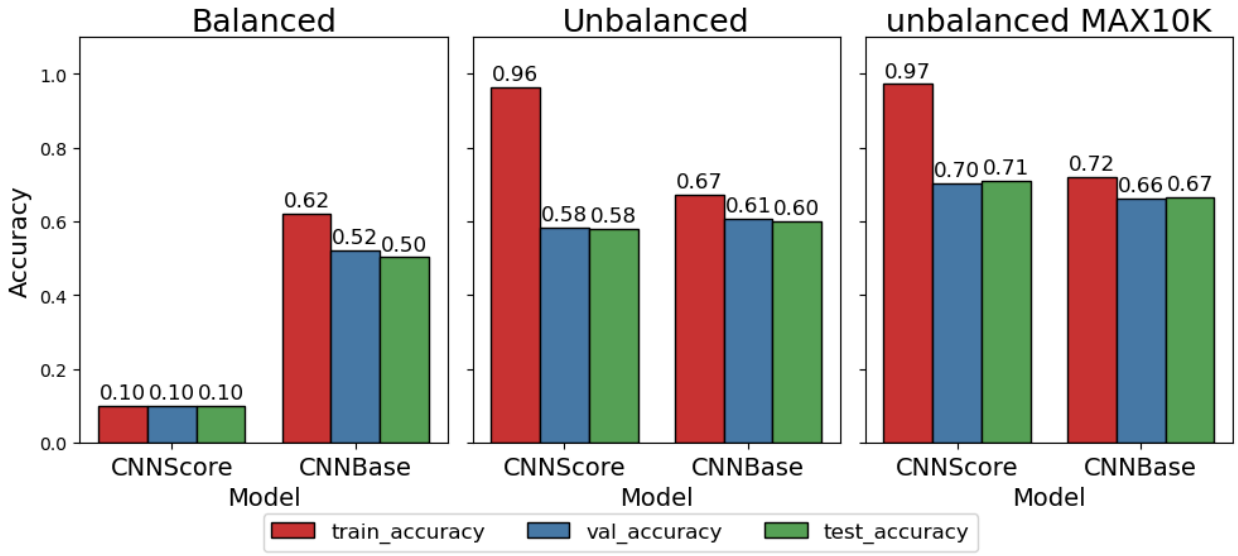


Figure 29. Comparison of dataset balancing strategies for RecOnly fragment classification. Performance evaluation across three preprocessing approaches: balanced, unbalanced, and unbalanced MAX10K using CNNBase and CNNScore architectures. Bar plots show train (red), validation (blue), and test (green) accuracies. The unbalanced MAX10K strategy demonstrates superior performance with better generalization, while the balanced strategy exhibits overfitting across models.

The evaluation of different dataset balancing strategies reveals distinct performance patterns and generalization capabilities across the tested approaches. The balanced strategy, which utilized the smallest number of samples per class, demonstrated clear signs of overfitting with CNNBase achieving a training accuracy of 0.62 but substantially lower validation (0.52) and test (0.50) accuracies, indicating poor generalization to unseen data. The standard unbalanced approach showed improved performance with CNNBase reaching more consistent accuracies across train (0.67), validation (0.61), and test (0.60) sets, suggesting better model stability. The unbalanced MAX10K strategy yielded the most promising results, with CNNBase achieving the highest and most balanced performance across all metrics (train: 0.72, validation: 0.66, test: 0.67), demonstrating effective generalization. The unbalanced MAX10K reveals severe overfitting with training accuracy reaching 0.97 compared to validation and test accuracies around 0.70-0.71 for

model with high complexity as CNNScore (Figure 28). The MAX10K sampling strategy's improvement maximize protein diversity by prioritizing different protein sequences before resampling from clusters. This also solve the class imbalance by taking similar proteins in cases were samples are missing. Therefore, maintaining richer structural representation while providing adequate sample sizes for effective learning.

5.2.3.3 Per-class accuracy and entropy analysis on different balancing strategies

The per-class analysis reveals substantial performance variation across fragment classes, with distinct patterns emerging between aromatic and aliphatic functional groups that differ significantly between balancing strategies. In the unbalanced dataset (Figure 29 A), aliphatic hydroxyl groups such as CO and amines attached to aromatic rings cN, showed reasonable performance with accuracy values of 0.7 and 0.67, respectively. Hydroxyl groups attached to aromatic rings (cO) showed the poorest performance (0.2 accuracy, 0.4 F1), along with amide-containing CC(=O)NC (0.3 accuracy, 0.4 F1) and simple aliphatic nitrogen-containing fragments CN and CN(C)C (both ~0.4 accuracy). The MAX 10K sampling strategy (Figure 28B) dramatically altered these performance hierarchies, with aromatic carbonyl c=O achieving the highest accuracy (0.8) and F1 score (0.9), while several previously high-performing aromatic fragments like cn(C)c dropped to moderate performance levels (0.6 accuracy). Notably, the MAX 10K strategy provided substantial improvements for previously poorly performing fragments with lower number of samples. with CN improving from 0.4 to 0.7 accuracy. This rebalancing effect demonstrates that the MAX 10K strategy not only addresses class imbalance but also reveals the difficulty of differencing similar chemical moieties.

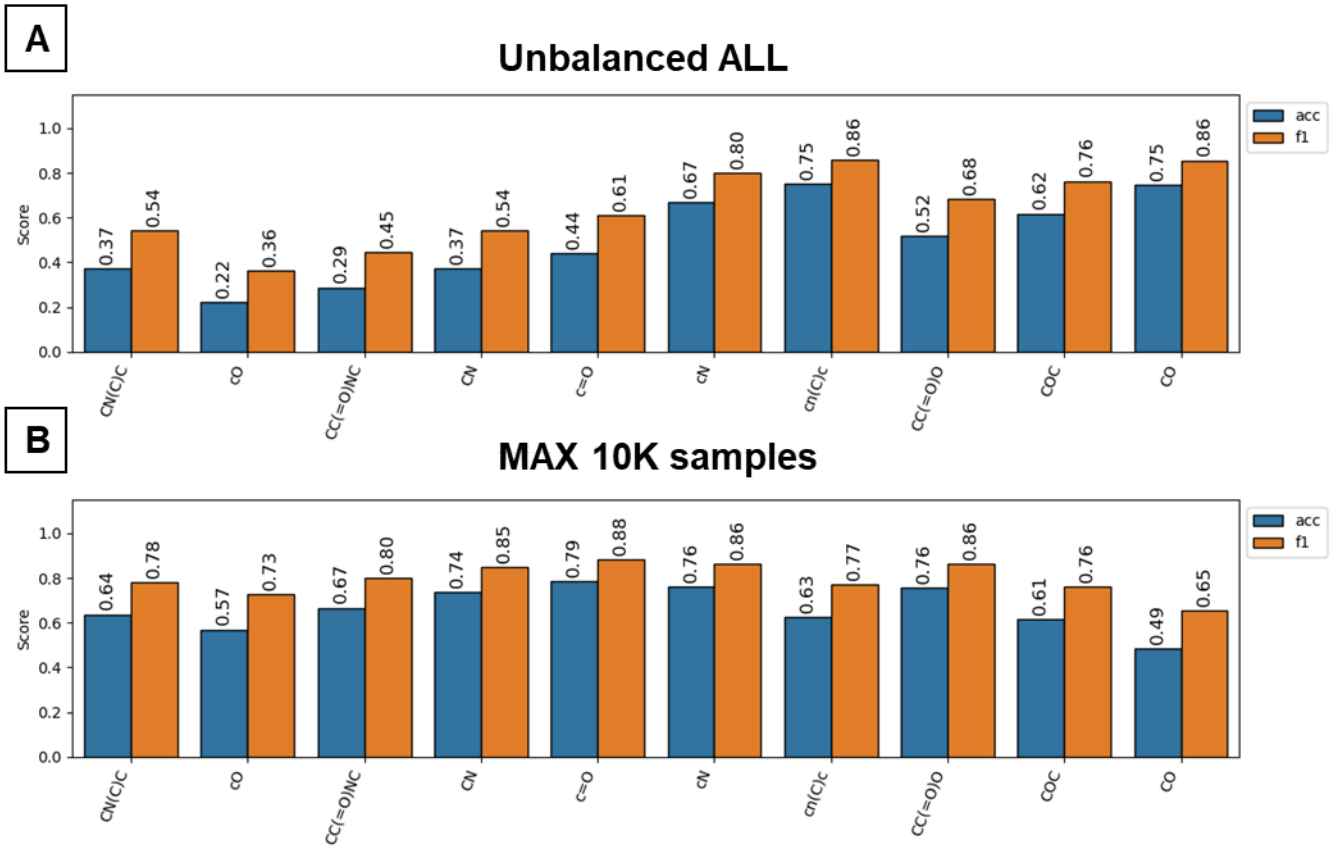


Figure 30Per-class performance comparison for RecOnly voxel configurations across balancing strategies. Bar plots showing accuracy (blue) and F1 scores (orange) for individual fragment classes using (A) unbalanced dataset with all available samples and (B) MAX 10K sampling strategy. Fragment classes include both aromatic (lowercase 'c') and aliphatic (uppercase 'C') representations. The MAX 10K strategy shows improved performance for several fragment types while maintaining chemical interpretability.

Entropy analysis demonstrates that model confidence correlates strongly with prediction accuracy in both datasets, with distinct improvements observed under the MAX 10K sampling strategy that reflect both reduced uncertainty and improved fragment-specific learning patterns. The unbalanced dataset (Figure 30 A) exhibited a broader entropy distribution with higher uncertainty peaks, where fragments such as CN(C)C displayed the highest average entropy (0.61), followed by amide-containing CC(=O)NC (0.58) and aromatic hydroxyl cO (0.43), indicating substantial model uncertainty in distinguishing these chemically similar moieties. Conversely, aromatic nitrogen-containing fragments demonstrated notably lower entropy values, with cN achieving 0.24 and cNCc reaching 0.32, suggesting that the model develops greater confidence when predicting these distinct heterocyclic features.

The MAX 10K strategy (Figure 29 B) produced systematically lower entropy values across most fragment classes, demonstrating effect of balanced sampling on model confidence, though the improvements varied significantly depending on fragment complexity and chemical distinctiveness. The most dramatic entropy reductions occurred for previously challenging fragments: CN(C)C entropy decreased from 0.61 to 0.36 (a 41% improvement), CC(=O)NC dropped from 0.58 to 0.25 (57% improvement), and cO improved from 0.43 to 0.33 (23% improvement), indicating that balanced training data enables the model to develop more discriminative features for these structurally complex groups. Fragments that already demonstrated low entropy in the unbalanced dataset maintained their confident prediction profiles under MAX 10K sampling, with cN decreasing only marginally from 0.24 to 0.22. Notably, two fragments exhibited entropy increases under MAX 10K sampling: COC rose from 0.38 to 0.42 and CO increased from 0.37 to 0.44, which may reflect the introduction of more structurally diverse protein environments that challenge the model's ability to identify consistent ether and hydroxyl binding patterns across the expanded protein sequence space.

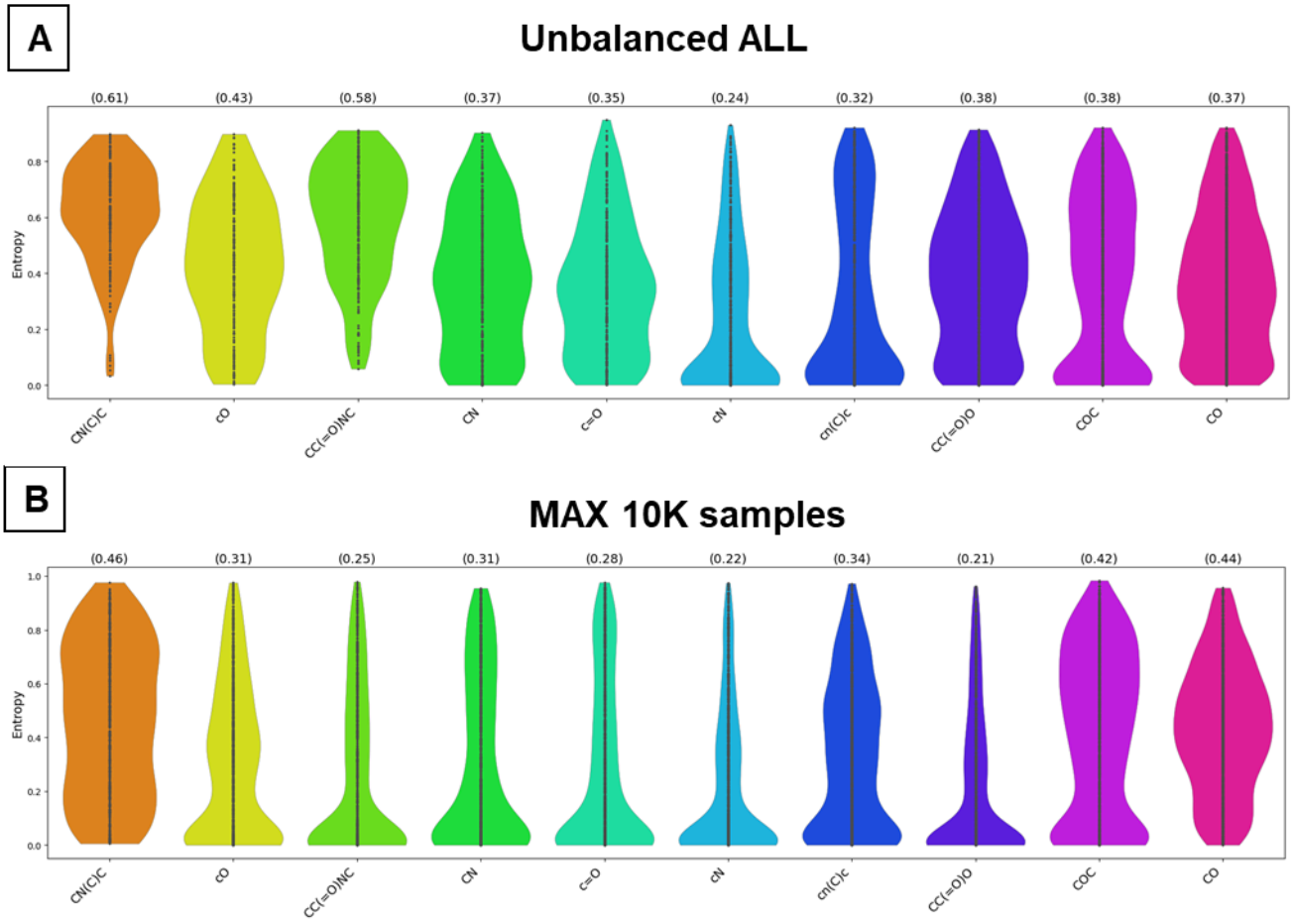


Figure 31. Per-class entropy distributions comparing balancing strategies for RecOnly fragment classification using CNNBase architecture. Violin plots showing entropy distributions for 10 representative fragment classes under (A) unbalanced dataset using all available samples and (B) MAX 10K sampling strategy. Average entropy values are shown in parentheses above each distribution, with lower values indicating higher model confidence.

5.2.3.4 Model parameters optimization :Effect of grid size and model complexity in model accuracy

Next, we evaluate the impact of grid size on model generalization performance across four CNN architectures using a fixed learning rate of 0.0001. While the grid size 10 Å produced the highest training accuracy (0.97 for CNNScore), this configuration exhibited significant overfitting, as evidenced by the substantial gap between training accuracy and test/validation performance (0.70 and 0.69, respectively). Grids sizes 12 Å and 6 Å showed better generalization with more balanced train-test-validation accuracy ratios (Figure 31). For practical fragment prediction, grids size 12 Å achieved the most reliable performance with test and validation accuracies ranging

from 0.63-0.67 across architectures, while grid size 6 Å showed slightly lower but consistent performance (0.62-0.66).

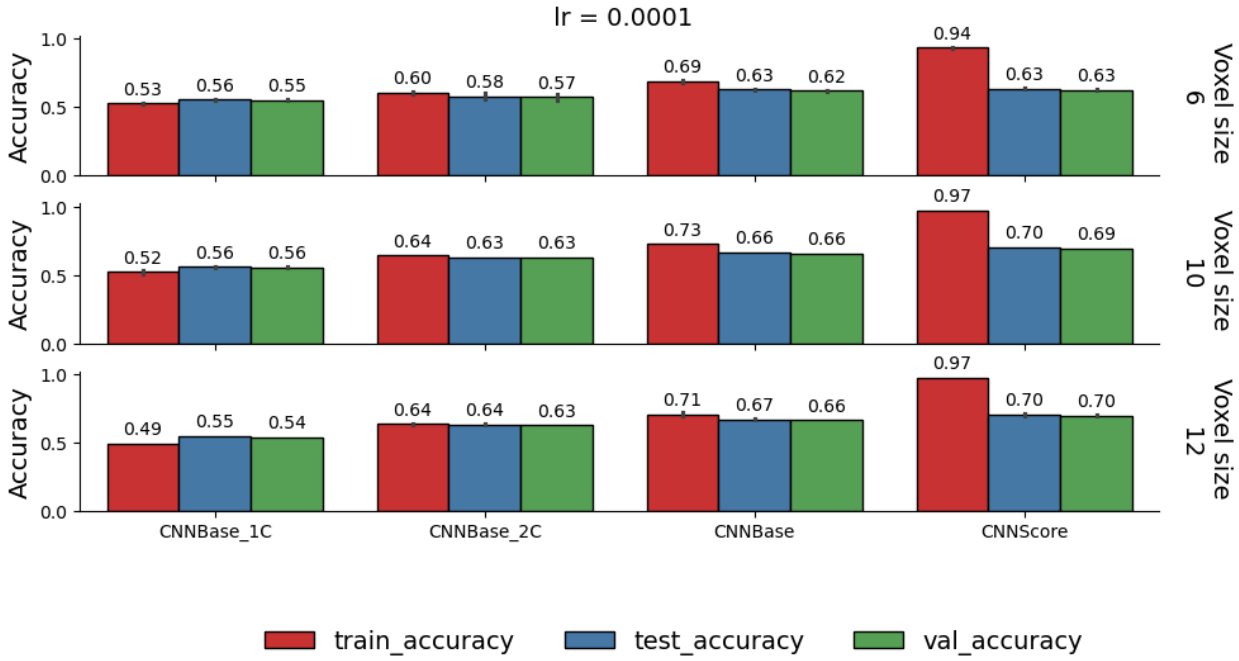


Figure 32. Effect of grids size on RecOnly model performance across CNN architectures. Comparison of train, test, and validation accuracies for four CNN architectures (CNNBase_1C, CNNBase_2C, CNNBase, CNNScore) using three different grids sizes (10 Å, 12 Å, 6 Å) with fixed learning rate of 0.0001. Grid size 10 Å shows overfitting with high training but moderate test/validation performance, while 12 Å provides better generalization balance.

5.2.3.5 Training parameter optimization: Learning rate and batch size effect on accuracy.

A comprehensive grid search was conducted to optimize the training parameters for the RecOnly models, evaluating learning rates from 1×10^{-5} to 1×10^{-3} and batch sizes of 32 and 64 across four CNN architectures. The analysis revealed that the learning rate was the most critical factor influencing model performance, while the batch size had a negligible impact. The results in Figure 31 showed a clear trend where lower, more conservative learning rates resulted in better generalization and avoided training instabilities. Both batch sizes tested provided adequate gradient estimates, indicating that the choice between them could be based on computational resources rather than performance.

The lowest learning rate of 1×10^{-5} consistently yielded the best and most stable results, with the CNNBase architecture providing the optimal balance between high accuracy (0.69) and controlled overfitting. In contrast, the standard rate of 1×10^{-4} caused the complex CNNScore model to suffer from overfitting, memorizing the training data (0.97 training accuracy) without improving generalization (0.71 test accuracy). The highest rate of 1×10^{-3} proved too aggressive, leading to complete training failure. Based on these findings, a learning rate of 1×10^{-5} with a batch size of 32 was selected as the optimal configuration, reinforcing the choice of the more robust CNNBase architecture for practical applications.

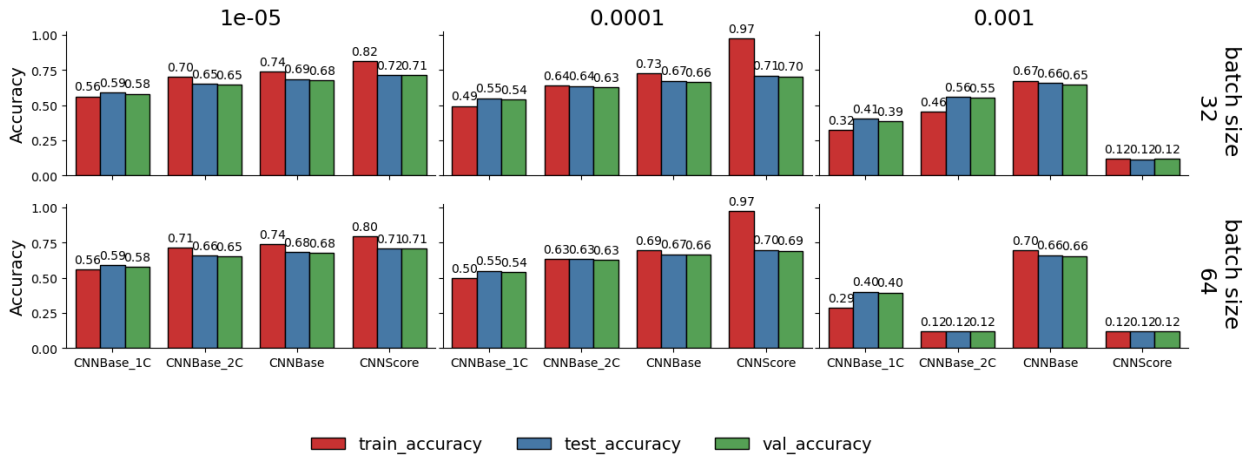


Figure 33 Impact of Learning Rate and Batch Size on Model Performance. Training, test, and validation accuracies for four different CNN architectures across three learning rates (1e-05, 0.0001, and 0.001) and two batch sizes (32 and 64). The results show that a conservative learning rate of 1e-05 yields the best generalization performance, while a high learning rate of 0.001 causes training instability. The choice of batch size has a negligible effect on the final accuracy.

5.2.3.6 Label Smoothing Fails to Mitigate Overfitting: Architecture Matters More Than Regularization

Then we explore the effectiveness of label smoothing as a regularization technique to address overfitting in RecOnly models using learning rate 1e-4. Label smoothing values of 0.0, 0.05, and 0.1 were tested across CNNBase and CNNScore architectures. CNNBase demonstrated consistent generalization performance across all label smoothing values, maintaining test and validation accuracies between 0.66-0.67 with minimal overfitting (training accuracies of 0.73-0.76). In contrast, CNNScore exhibited severe overfitting regardless of label smoothing implementation, with training accuracies reaching 0.97, while test and validation performance

remained stable at 0.70-0.71. Surprisingly, label smoothing failed to mitigate overfitting in CNNScore and even slightly worsened the training-validation gap at higher smoothing values (0.05 and 0.1), suggesting that architectural complexity rather than label confidence is the primary driver of overfitting in this model. (Figure 32)

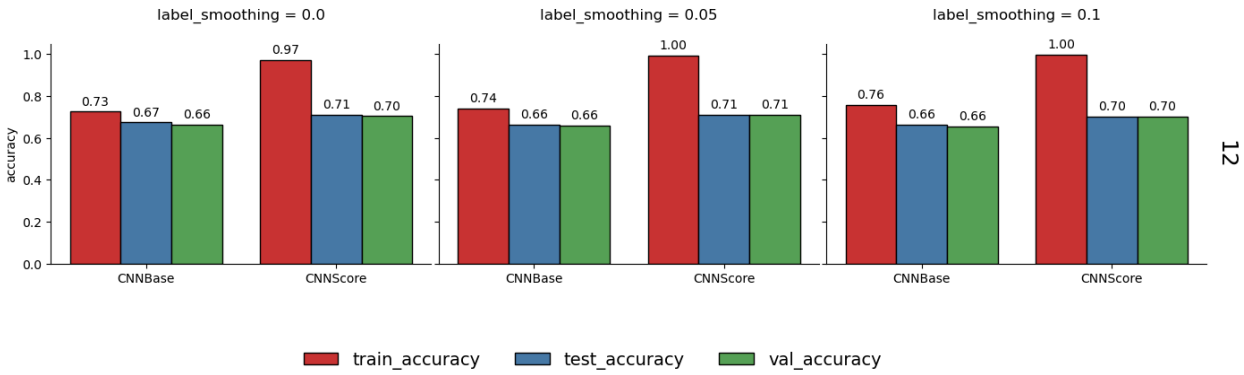


Figure 34. Effect of label smoothing on RecOnly model performance using grids size 12 Å and learning rate 1e-4. Comparison of train, test, and validation accuracies for CNNBase and CNNScore architectures across three label smoothing values (0.0, 0.05, 0.1). Label smoothing does not effectively reduce overfitting in CNNScore, while CNNBase maintains consistent generalization performance across all smoothing values.

5.2.3.7 Multi-Class Classification Performance and Confidence Analysis

We explore the impact of classification complexity on model performance and generalization (Figure 34). As the number of fragment classes increased from 10 to 60, both architectures showed expected decreases in absolute accuracy but improved generalization characteristics. At 10 classes, CNNScore exhibited substantial overfitting (0.99 training vs. 0.71 test accuracy), while CNNBase maintained better balance (0.74 training vs. 0.67 test accuracy). The 30-class configuration provided a reasonable compromise with CNNBase achieving balanced performance (0.53-0.54 across all sets) and CNNScore showing reduced overfitting (0.94 training vs. 0.61-0.62 test/validation). The 60-class setting further improved generalization, with CNNBase achieving near-perfect balance (0.49-0.51) and CNNScore showing the best generalization performance observed (0.91 training vs. 0.58 test/validation). These results indicate that increasing classification complexity naturally regularizes model performance and

that CNNBase architecture provides superior generalization across all tested configurations, making it the preferred choice for practical fragment prediction applications.

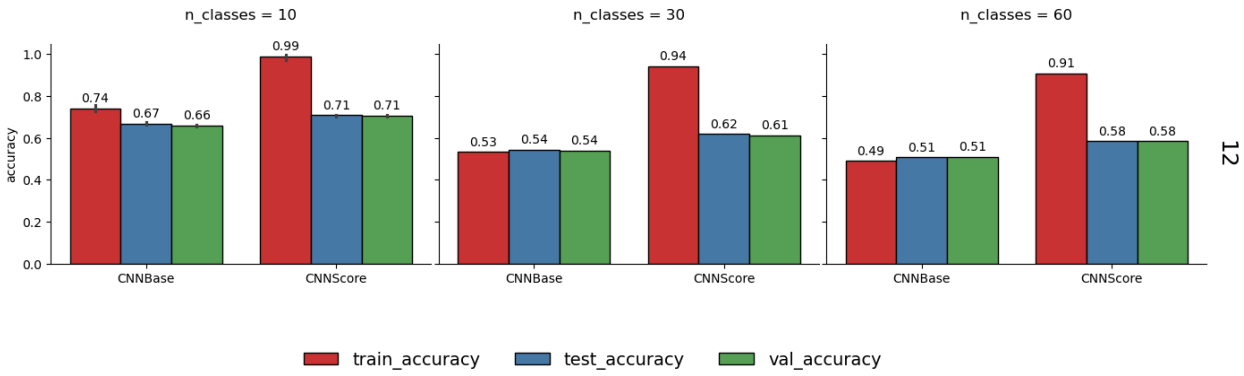


Figure 35. Multi-class fragment classification performance using grid size 12 Å and learning rate 1e-4. Performance comparison across different numbers of fragment classes (10, 30, 60) for CNNBase and CNNScore architectures. Increasing the number of classes reduces absolute accuracy but improves model generalization, with CNNBase consistently demonstrating better balance between training and test performance than CNNScore.

The top-K accuracy analysis reveals important insights into model performance beyond standard single-prediction metrics, demonstrating substantial improvements when considering multiple predictions. Figure 35A shows that both 30-class and 60-class models achieve significant performance gains when evaluating top-K predictions, with top-5 accuracy reaching 0.8 and 0.8 compared to top-1 accuracies of 0.5 and 0.5, respectively. This 30-35% improvement indicates that correct fragment predictions are frequently among the model's top candidates, suggesting that the learned feature representations capture meaningful chemical relationships even when the highest-confidence prediction is incorrect. The 30-class model consistently outperforms the 60-class model across all top-K values, reflecting the expected trade-off between fragment library size and classification accuracy. Per-fragment analysis (Figure 35 B-C) reveals significant heterogeneity in fragment classification difficulty, with some fragments achieving near-perfect top-1 accuracy while others require top-5 considerations to reach acceptable performance levels. In the 30-class model, high-performing fragments like CS(=O)(=O)O and aromatic systems show top-1 accuracies above 0.8, while challenging fragments such as thiourea derivatives S=C([#7][#6])[#7][#6] and guanidine-like structures exhibit top-1 accuracies below 0.2 but

recover to 0.4-0.6 in top-5 evaluations. The 60-class model shows more pronounced performance variation, with top-performing fragments maintaining accuracies above 0.8 and difficult fragments requiring top-3 or top-5 considerations for reliable prediction. This analysis demonstrates that practical fragment suggestion systems should implement ranked prediction outputs rather than single classifications, as the top-3 to top-5 predictions provide substantially more useful information for medicinal chemistry applications.

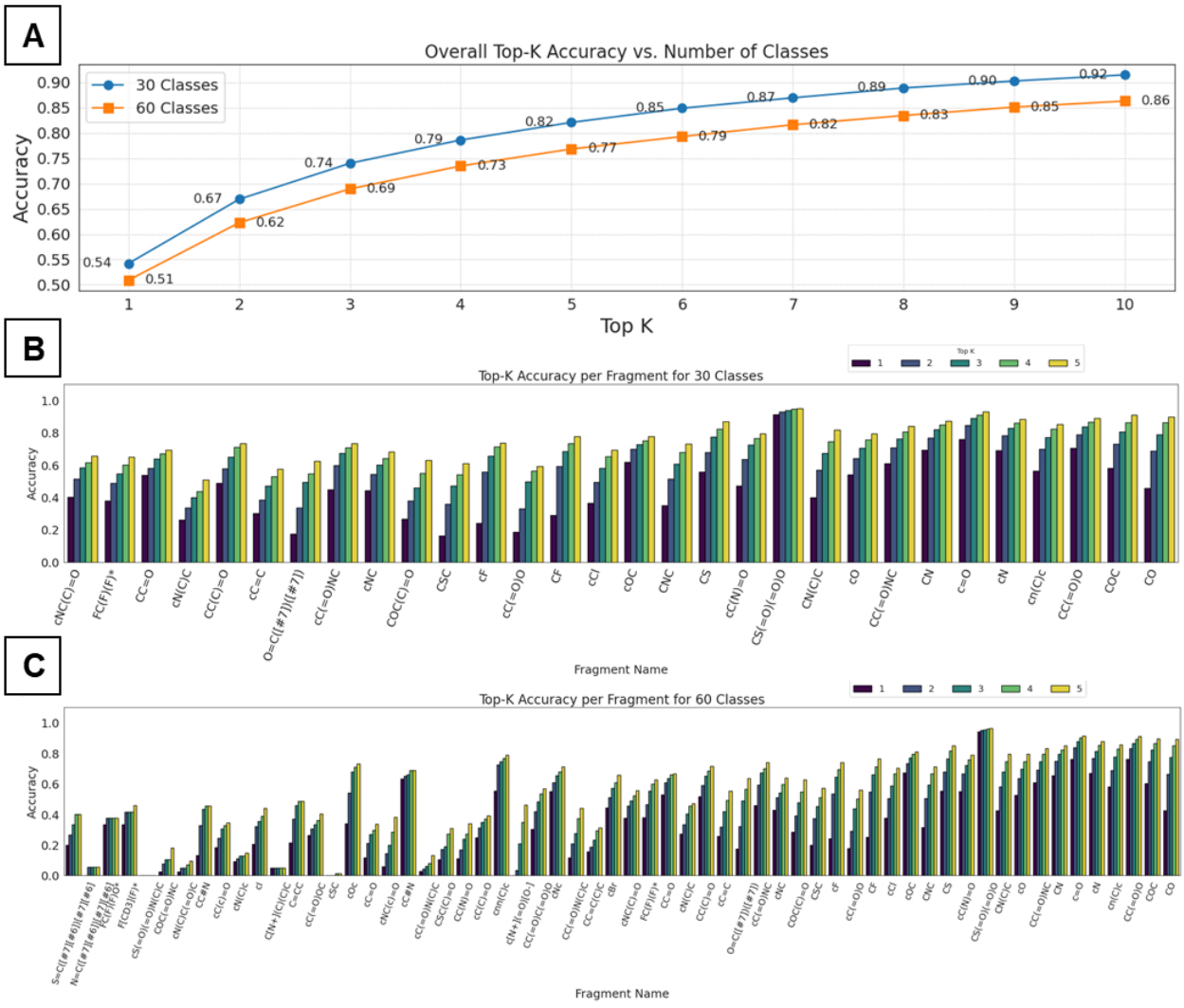


Figure 36 Top-K accuracy analysis for multi-class fragment classification using CNNBase architecture. (A) Overall top-K accuracy comparison between 30-class and 60-class models, showing cumulative performance improvement as K increases from 1 to 10. (B) Per-fragment top-K accuracy distribution for 30-class model, displaying individual fragment performance across different K values (1-5). (C) Per-fragment top-K accuracy distribution for 60-class model.

5.2.3.7.1 Sequence-Based Splitting Validates Generalization to Novel Protein Families

To ensure our performance metrics were robust and representative of real-world applications, we evaluated the impact of the data splitting methodology and the inclusion of more fragments to predict. Because random splitting distributes data uniformly, a sequence-based split (e.g., at a 50% identity threshold) provides an additional test by preventing information leakage from homologous proteins between the training and test sets. Here, CNNBase maintains consistent generalization performance across both splitting methodologies and varying class complexities. The sequence-based split, which provides a more stringent evaluation by ensuring no protein homologs appear in both training and test sets, achieved comparable performance to random splitting across all accuracy metrics (0.70 val accuracy for 10 model classes). This consistency indicates that the model learns transferable fragment-binding site relationships rather than memorizing protein-specific patterns. The minimal performance difference between random and sequence-based splitting (typically <0.02 accuracy difference) suggests robust feature learning that generalizes well to unseen protein families. These results validate that the classification approach can handle realistic fragment library sizes while maintaining reliable performance on completely novel protein-ligand systems, supporting its applicability for practical structure-based drug design applications.

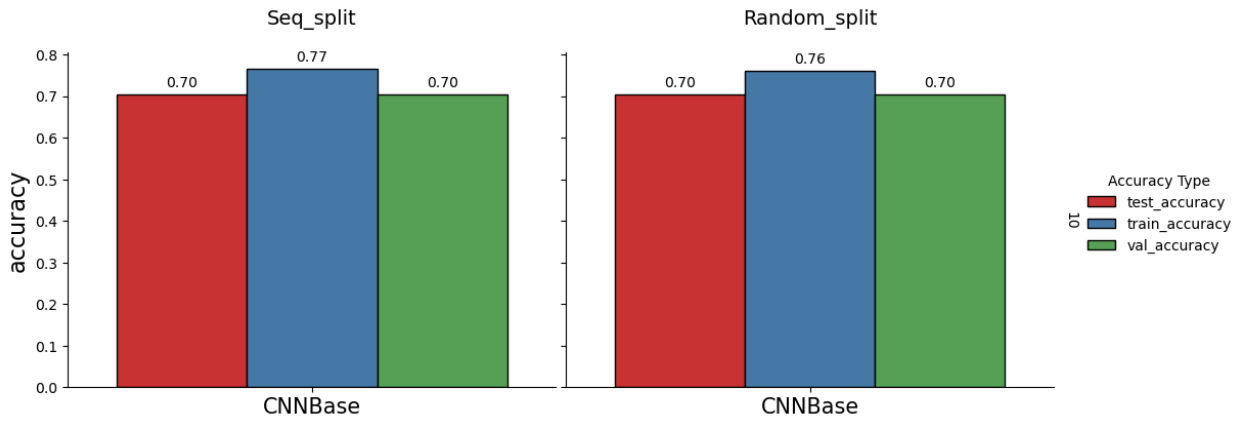


Figure 37. Impact of data splitting methodology and fragment library size on CNNBase performance**.** Comparison of train, test, and validation accuracies using sequence-based split (left panels) versus random split (right panels) for 17 10-class fragment classification models. Both

splitting strategies achieve similar performance levels, demonstrating robust generalization to unseen protein families and consistent performance across different fragment library sizes.

5.2.4 Chemistry-Informed Fragment Library Optimization

The comprehensive hyperparameter optimization revealed an intricate trade-off between classification accuracy and the number of fragment classes, where increasing from 10 to 60 classes naturally regularized model performance but reduced absolute accuracy (from 0.67 to 0.58 test accuracy for CNNBase). This systematic exploration—encompassing data preprocessing strategies, voxel smoothing parameters (r=1.75 providing optimal performance), balancing approaches (MAX10K showing superior generalization), and model architectures (CNNBase demonstrating consistent stability)—ultimately pointed to a fundamental insight: the model's performance limitations were not solely technical but also reflected the chemical redundancy in our fragment representation. The per-class performance analysis revealed that chemically equivalent fragments, differing only in their attachment to aromatic versus aliphatic carbons (e.g., cO vs CO, cN vs CN), exhibited significantly different classification accuracies despite their similar chemical profiles. This observation, combined with the top-K accuracy analysis showing that correct fragments frequently appeared within the top-5 predictions (80% accuracy), suggested that the model was learning meaningful chemical relationships but was being penalized by artificially granular class distinctions.

To address this, we implemented a chemistry-informed merging strategy that consolidated equivalent functional groups regardless of their aromatic or aliphatic carbon attachments, effectively reducing redundant classes while preserving chemical diversity. Furthermore, recognizing that the Ertl fragmentation algorithm's focus on heteroatom-containing groups missed important structural motifs used in medicinal chemistry, we complemented our fragment library with full aromatic ring systems identified by RDKit's Murcko scaffold splitter algorithm. These aromatic scaffolds—including common medicinal chemistry building blocks like substituted benzenes, pyridines, and fused ring systems—represent critical pharmacophores that were previously absent from our functional group-centric approach.

This hybrid Ertl_MedChem+Murcko_MedChem strategy yielded a refined library of 30 fragment classes, each with at least 1500 samples to ensure robust training (Figure 37). The final

fragment set represents a carefully balanced chemical space that includes: (i) merged functional groups where aromatic and aliphatic variants are combined (e.g., CO/cO hydroxyl groups, CN/cN amines, COC/cOc ethers), reflecting their similar interaction profiles in medicinal chemistry; (ii) distinct amide-containing fragments maintained as a separate class due to their unique binding characteristics (CC(=O)N, cC(=O)N variants with different substitution patterns); (iii) essential aromatic scaffolds including six-membered rings (benzene, pyridine), five-membered heterocycles (pyrrole, thiophene, furan), and bicyclic systems that serve as common pharmacophoric elements; and (iv) specialized functional groups critical for drug-target interactions such as sulfonamides (CS), halogen substituents (CCl, CF), and charged moieties.

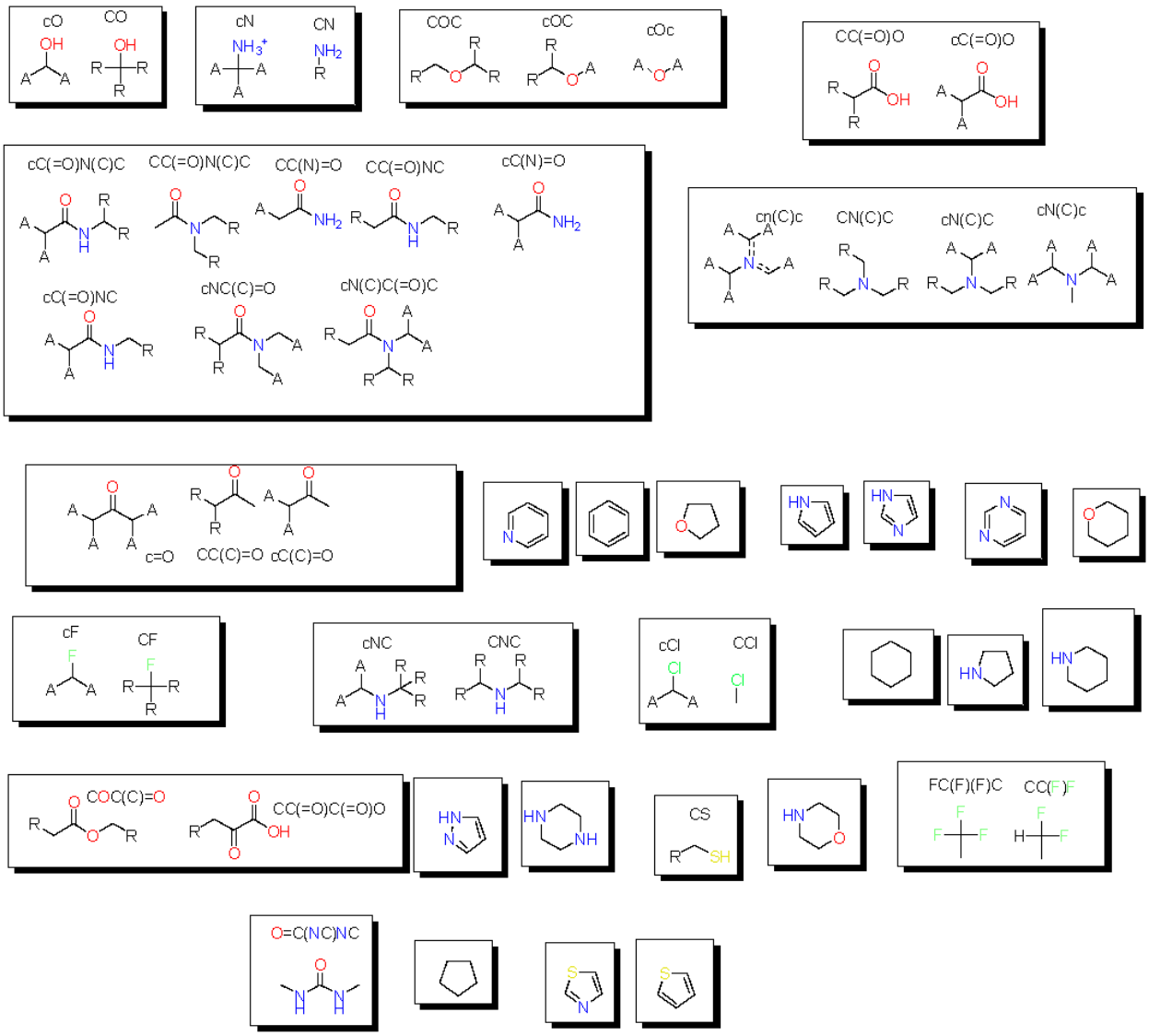


Figure 38. Curated fragment library for classification-based lead optimization. The 30 fragment classes selected for the hybrid Ertl_MedChem+Murcko_MedChem model, each containing at least 1,500 protein-ligand complex samples after MAX10K sampling. Fragments are organized by chemical class and include: (top row) merged hydroxyl, amine, and ether groups combining aromatic and aliphatic variants; (second row) amide-containing and amine fragments with varying substitution patterns maintained as distinct classes due to unique binding profiles; (third row) carbonyl-containing groups and essential aromatic scaffolds including six-membered rings (benzene, pyridine) and five-membered heterocycles (pyrrole, thiophene, furan); (bottom rows) halogenated fragments, sulfonamide groups, bicyclic systems, and specialized heterocycles commonly employed in medicinal chemistry.

The classification results reveal a clear performance hierarchy that correlates strongly with fragment complexity and chemical distinctiveness. Simple aromatic scaffolds demonstrate the highest classification accuracy, with pyridine (C1=CN=CC=C1) achieving 0.79 accuracy and benzene (C1=CC=CC=C1) reaching 0.70, suggesting that these rigid, planar structures create distinctive binding site signatures that the model can readily identify. In contrast, the most complex merged class, combining nine different amide-related fragments shows moderate performance at 0.57 accuracy despite representing the largest chemical diversity, indicating that our merging strategy successfully grouped chemically similar fragments that the model struggles to distinguish individually.

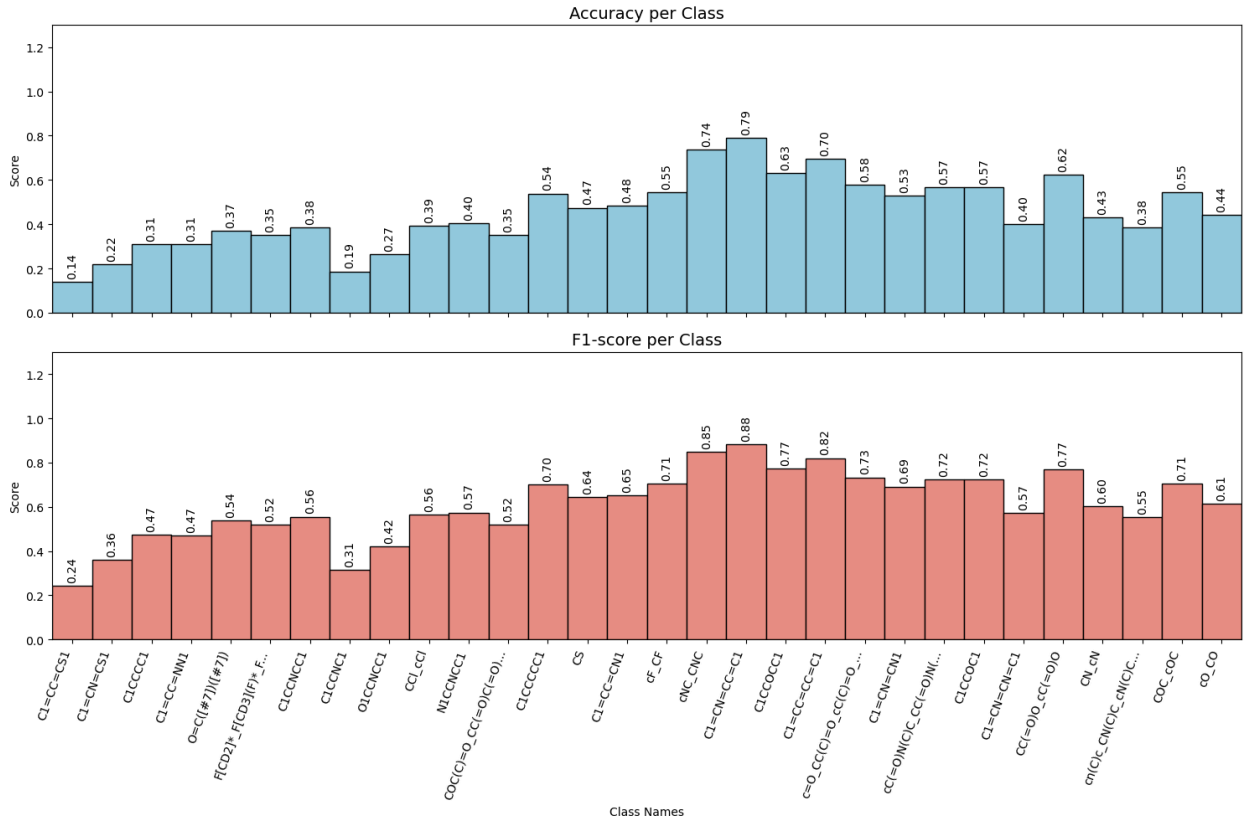


Figure 39. Per-class performance metrics for the final 30-fragment classification model. (Top) Accuracy scores for each fragment class, ranging from 0.14 to 0.79, with aromatic scaffolds (C1=CN=CC=C1, C1=CC=CC=C1) and merged amine groups (cNC_CNC) showing the highest performance. (Bottom) F1-scores demonstrating class-balanced performance, with values ranging from 0.24 to 0.88.

The merged functional groups exhibit intermediate performance that justifies our consolidation approach: the combined amine class (cNC_CNC) achieves remarkably high accuracy (0.74), while merged ethers (COC_cOC) and hydroxyls (cO_CO) show moderate performance (0.55 and 0.44 respectively), suggesting these merged classes capture meaningful chemical similarities while maintaining discriminative power. Notably, halogenated fragments demonstrate differential performance with chlorides (CCl_cCl) at 0.39 accuracy versus fluorides (cF_CF) at 0.55, reflecting the distinct electronic and steric effects of different halogens on binding site interactions. The entropy analysis reinforces these findings, showing that high-performing fragments like aromatic heterocycles exhibit narrow, low-entropy distributions indicative of confident predictions, while poorly performing classes display broader entropy spreads reflecting model uncertainty (See Figure 39. This performance gradient validates our hybrid fragmentation

approach: fragments that are chemically distinct and rigid (aromatic scaffolds) are learned effectively, while those requiring subtle discrimination within merged classes present appropriate classification challenges that align with their chemical similarity, ultimately producing a model that balances computational accuracy with medicinal chemistry intuition.

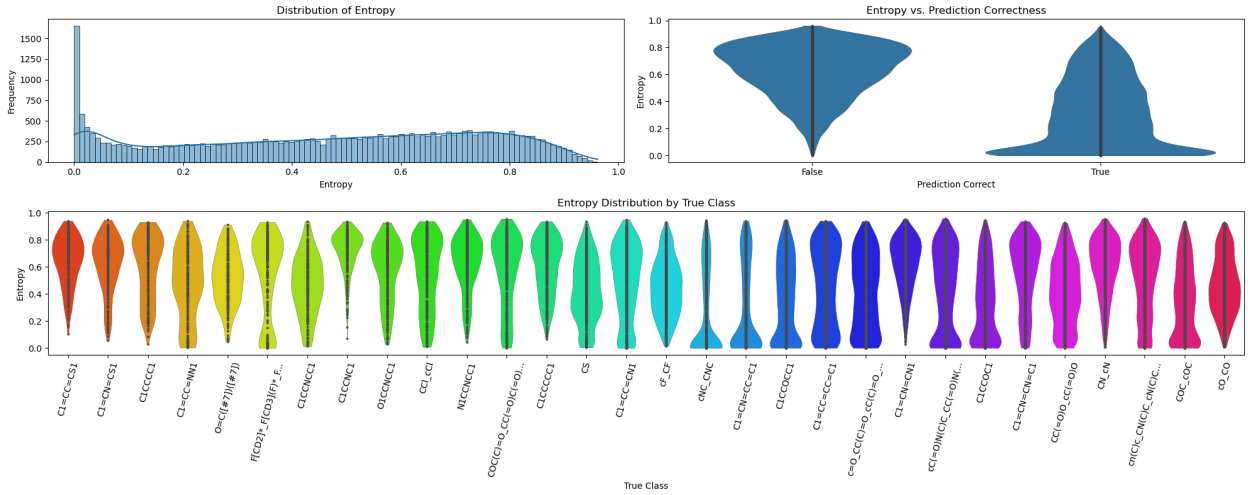


Figure 40. Entropy analysis and prediction confidence for the 30-fragment classification model. (Top left) Overall entropy distribution showing a sharp peak near zero with a long tail extending to 1.0, indicating predominantly confident predictions with some uncertain cases. (Top right) Violin plots comparing entropy distributions between incorrect` (False) and correct (True) predictions, demonstrating clear separation with correct predictions exhibiting lower entropy (higher confidence) and incorrect predictions showing broader entropy distributions. (Bottom) Per-fragment entropy distributions showing class-specific prediction confidence.

5.2.5 Protease, Kinases and Polymerase inhibitors as validation cases.

5.2.5.1 SARS-CoV-2 3CL Protease Inhibitors

To evaluate the performance of our CNN-based fragment prediction model in a real-world drug discovery context, we selected two structurally related SARS-CoV-2 3CL protease inhibitors. The validation case examines compounds 7XB and 7YY, produced in the structure-based optimization campaign that ultimately yielded a 90-fold improvement in enzymatic inhibitory activity. The initial hit compound 7XB (IC50 = 8.6 μM) features a central triazinedione scaffold decorated with a trifluorobenzyl moiety and a difluoromethoxy-2-methylbenzene substituent, discovered through virtual screening of an in-house compound library.[168] Through systematic structural optimization guided by X-ray crystallography, the difluoromethoxybenzene group was

replaced with a 6-chloro-2-methyl-2H-indazole moiety to generate lead compound 7YY (IC50 = 0.096 μM), which maintains the core triazinedione-trifluorobenzyl architecture while introducing an expanded heterocyclic system. This structural evolution from a relatively simple aromatic substituent to a complex fused heterocycle provides an ideal test case for assessing whether our fragment prediction model can accurately identify both conserved pharmacophoric elements and distinguish subtle chemical modifications that drive potency improvements in lead optimization. The following analysis examines how well the model captures these structural features and whether its predictions align with the medicinal chemistry rationale that guided this successful drug discovery program.

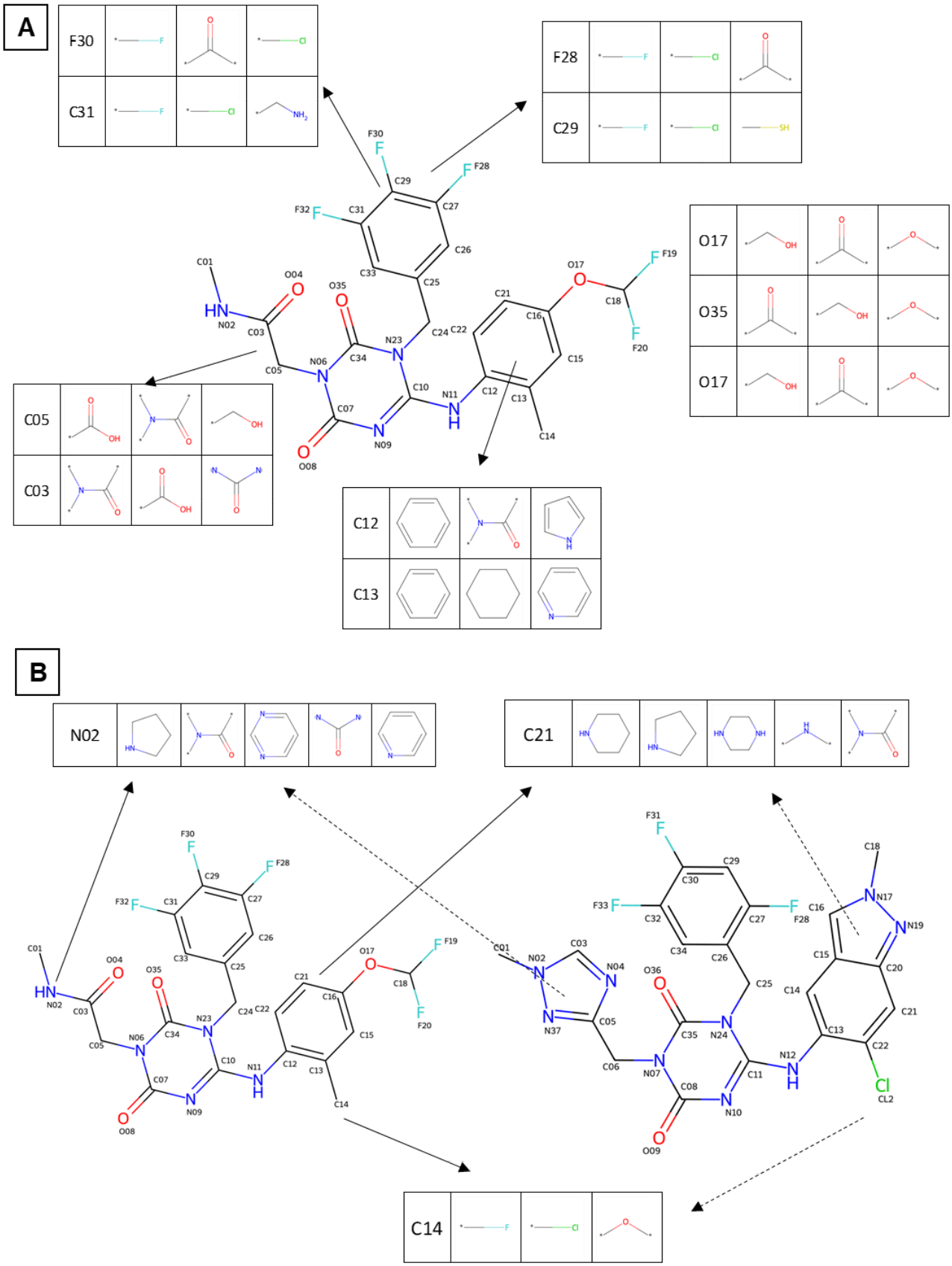

A
F30
C31
F28
C29
O17
O35
O17
C05
C03
C12
C13
B
N02
C21
C14

Figure 41. Fragment prediction analysis for SARS-CoV-2 3CLpro inhibitors demonstrating model accuracy and lead optimization. (A) Hit compound 7XB (IC50 = 8.6 μM) with CNNBase model fragment predictions for key structural elements. Boxes display top predicted fragments for selected atoms, with successful predictions highlighted for the triazinedione core (C03, C05), trifluorobenzyl moiety (F28, F30, F32, C29, C31), difluoromethoxy group (O17, O35), and aromatic regions (C12, C13). (B) Optimized lead compound 7YY (IC50 = 0.096 μM) showing fragment predictions that capture the structural evolution, including the added chloroindazole moiety (C14, with Cl2 chlorine substituent) and expanded heterocyclic framework (N02, C21). Fragment structures are shown in order of prediction confidence from left to right

The CNNBase model's accurate fragment predictions for the hit compound 7XB, showed in Figure 40-A highlights key structural elements where the model correctly identified chemical moieties. The central triazinedione scaffold shows precise predictions for atoms C03 and C05, with the model correctly identifying amide-containing fragments. The trifluorobenzyl substituent predictions demonstrate excellent accuracy, with F28, F30, F32,C29 and C31 correctly assigned fluorine-containing fragments. The difluoromethoxy group is well-characterized through atoms O17 and O35, where the model successfully predicts hydroxyl (*OH) and ether (O) fragments that reflect the oxygen functionalities. Notably, the aromatic regions C12 and C13 show strong predictions of benzene ring systems and pyridine-like heterocycles, demonstrating the model's ability to distinguish between different aromatic scaffolds within the same molecular framework.

Having the structure of the optimized lead compound 7YY, allow us to identify how the model captures structural modifications that enhance potency from the initial hit. The expanded heterocyclic system is evident in the predictions for N02, which shows pyrrolidine and piperidine-like fragments alongside aromatic systems, reflecting the increased complexity of the optimized scaffold. The chloroindazole moiety, a key addition in the lead optimization, is represented through atom C14's predictions showing fluorine and chlorine-containing fragments, demonstrating the model's recognition of halogen substitution patterns critical for improved binding affinity. The predictions for C21 reveal interesting fragment suggestions including piperazine-like rings and amine-containing structures, indicating the model's interpretation of the extended nitrogen-rich heterocyclic framework.

5.2.3.2 ERK2 kinase inhibitors

The following analysis examines the performance of our CNN-based fragment prediction model on a series of ERK2 kinase inhibitors. These compounds, ranging from the initial hit (compound 8X2) to highly optimized lead molecules (compounds 8X5-8XH), showcase a systematic progression of structural modifications aimed at improving potency, selectivity, and drug-like properties.[169] The series demonstrates key medicinal chemistry strategies including the addition of benzyl groups to access the glycine-rich loop, replacement with methoxyethyl groups to reduce lipophilicity and the introduction of pyrazole moieties. By analyzing the model's fragment predictions across these structurally related yet functionally diverse molecules, we can assess its ability to recognize conserved fragments (the pyrimidine-pyrrololactam core), distinguish between different substituents (benzyl versus methoxyethyl), and identify subtle stereochemical variations that significantly impact binding affinity.

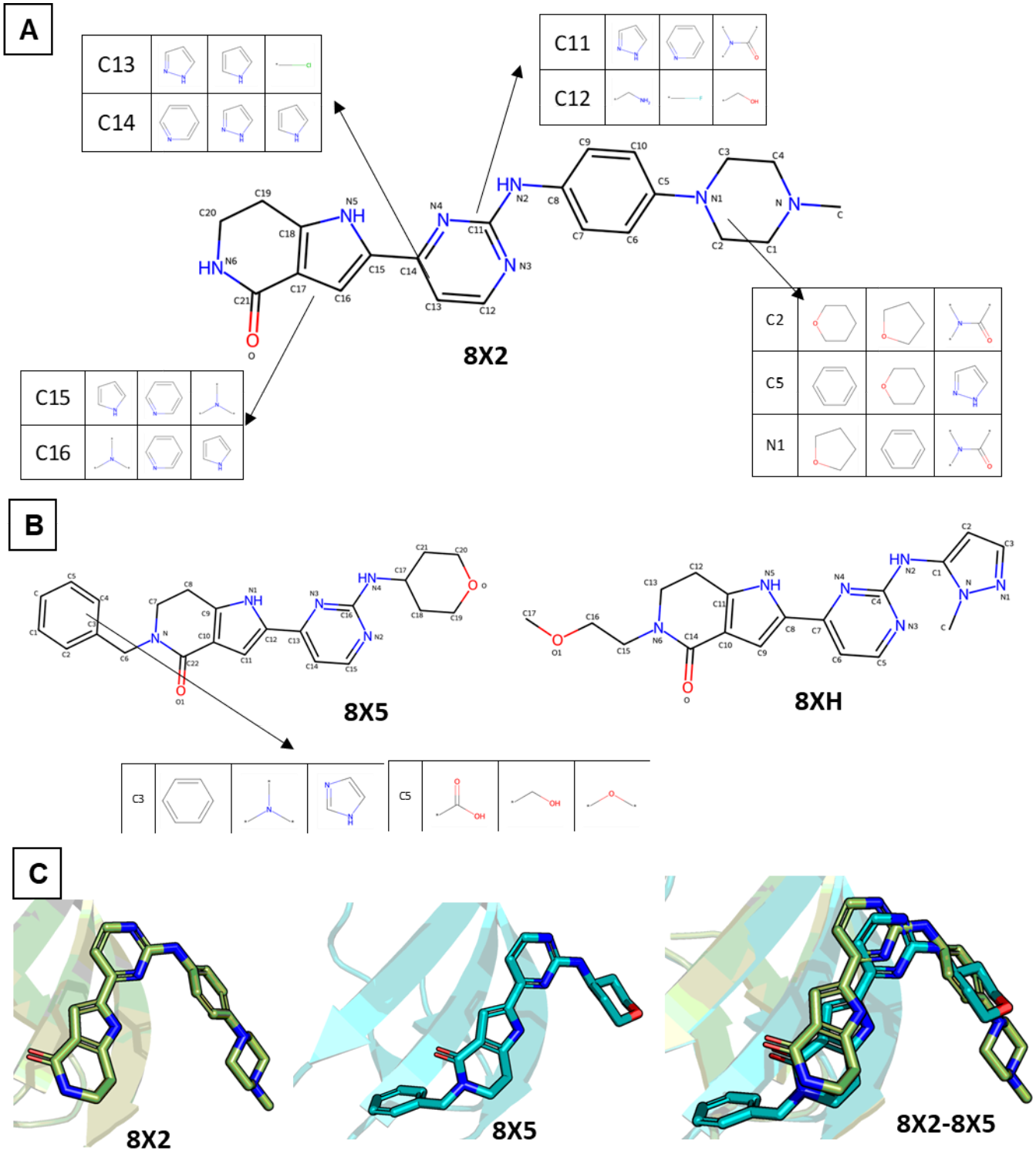


Figure 42. CNN-based fragment prediction analysis of ERK2 kinase inhibitors during lead optimization. (A) Fragment predictions for compound 8X2, the initial hit compound, showing predicted fragments for key structural regions. Each box displays the top 3 predicted fragments for the corresponding atomic center, with fragment classes including aromatic heterocycles (C11,

C12: pyrimidine variants), cyclic ethers (C2, C5: morpholine-related), and nitrogen based heterocycles groups (C15, C16). (B) Comparative fragment predictions for the benzyl-modified analog 8X5 and methoxyethyl derivative 8XH, highlighting how structural modifications affect model predictions. Fragment boxes demonstrate the model's ability to distinguish between different substituent patterns while maintaining recognition of the conserved core pharmacophore. (C) Crystal structure overlays showing the binding modes of 8X2 (green) and 8X5 (cyan) in the ERK2 active site, with the superimposed structures (8X2-8X5) illustrating the spatial relationships that influence fragment predictions.

The fragment prediction analysis of compound 8X2 showed the model's robust recognition of the core heterocyclic framework essential for ERK2 binding. The pyrimidine core atoms, C12 and C13, receive highly accurate predictions of nitrogen-rich heterocyclic fragments. The morpholine substituent is exceptionally well-characterized, with the model consistently predicting six-membered oxygen-containing ring fragments that accurately reflect the ether functionality within the saturated ring system. Note that the position of N1 atom in 8X2 aligns perfectly with the morphine ring (Figure 41-C). Additionally, the imidazole ring was well characterized by nitrogen-rich heterocyclic fragments.

Compound 8X5, featuring the potency-enhancing benzyl group that accesses the glycine-rich loop, showcases excellent model performance in aromatic recognition while highlighting challenges in specific substitution pattern identification. The newly introduced benzyl moiety receives strong and consistent aromatic predictions (C1=CC=CC=C1) at the atom C3. Interestingly, the prediction of the atom C5 corresponds to oxygen-rich fragments, feature that correlates with the 8XH structure and the methoxyethyl substitution that increase binding.

5.2.3.3 HCV NS5B Polymerase Inhibitors

To further assess our model's capabilities in a lead optimization scenario involving complex linker modifications, we selected a case study on non-nucleoside inhibitors of the Hepatitis C Virus (HCV) NS5B polymerase, an essential enzyme for viral replication. The study by Schoenfeld et al.[170] details the structure-based optimization of a fragment-derived hit into a clinical candidate, providing an excellent real-world test for our fragment prediction model.

This validation case focuses on two key analogues, 28Q and 28R, which represent distinct strategies for connecting a pyridone-based core to a distal N-arylmethanesulfonamide moiety.

Compound 4 employs a simple, flexible ethylene linker, whereas compound 12 utilizes a rigid, fused oxazole ring system designed to conformationally restrict the molecule into its active binding pose. This pair of inhibitors provides an ideal challenge for our CNN-based model. The analysis will test whether the model can not only identify the conserved core fragments but also differentiate between the subtle environmental changes required to accommodate a flexible versus a rigid linker. We will evaluate if the model's predictions for the linker region align with the observed medicinal chemistry outcomes, assessing its potential to guide complex linker design—a critical and often challenging task in structure-based drug discovery.

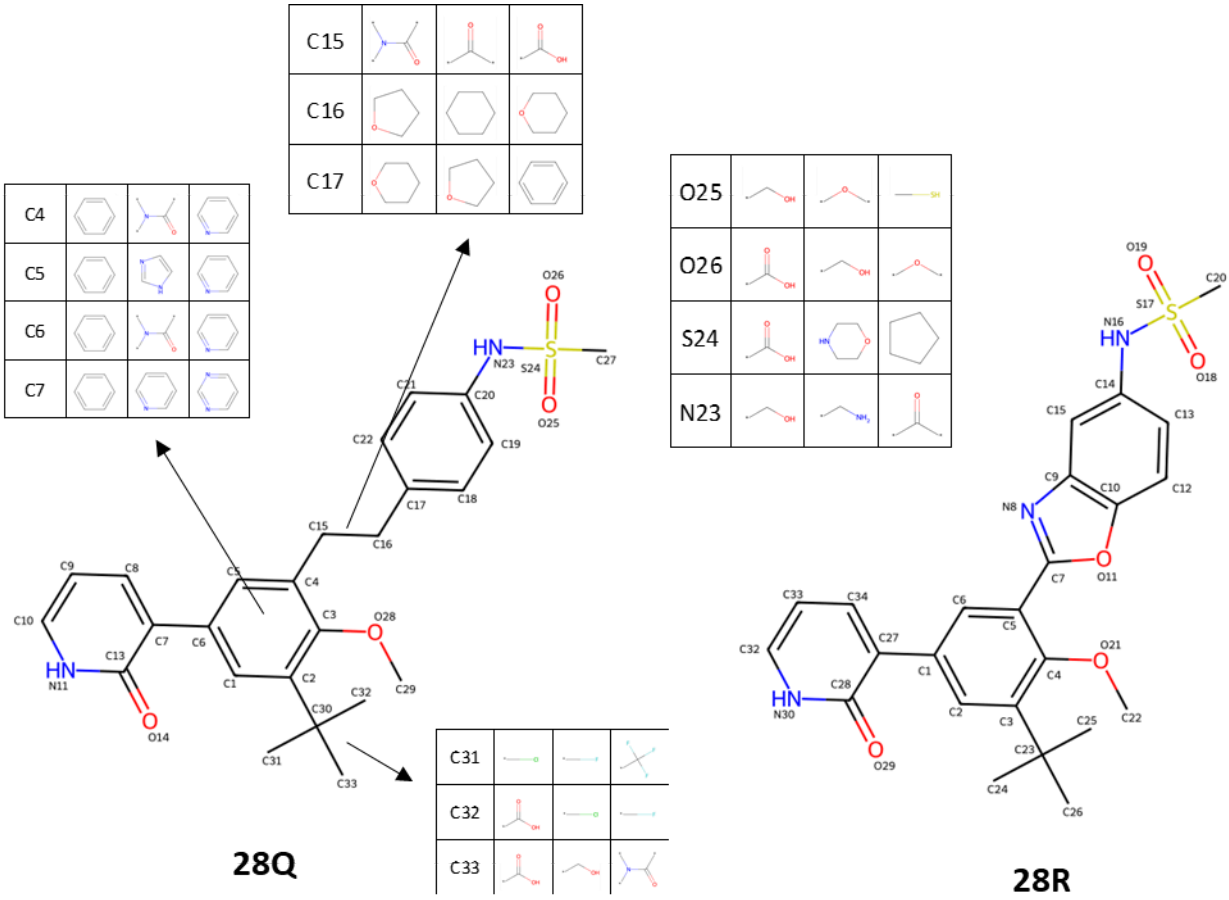


Figure 43. Compound 28Q featuring a flexible ethylene linker, showing fragment predictions for key structural regions including the aromatic core (C4-C7), the linker region (C15-C17), the sulfonamide moiety (N23, S24, O25, O26), and the tert-butyl group (C31-C33). Each box displays the top 3 predicted fragments for the corresponding atomic center, demonstrating the model's difficulty in accurately identifying aliphatic regions and essential functional groups. Compound 28R with the conformationally restricted oxazole ring system, illustrating the structural evolution from flexible to rigid linker design.

The model's performance on ligand 28Q, the inhibitor with the flexible ethylene linker, is largely inaccurate, demonstrating significant difficulty in identifying most of the key functional groups. The predictions for the core aromatic ring (atoms C1-C6) are correctly identified as belonging to a benzene ring. The model completely fails to recognize the aliphatic ethylene linker (C7, C8) and the tert-butyl group (C30-C33), predicting aromatic fragments or incorrect functional groups for these saturated carbons. Similarly, the methoxy group is misidentified, with predictions for O28 and C29 pointing to amides or heterocyclic rings. The most significant failure is the complete misidentification of the sulfonamide moiety; the model provides incorrect and generic predictions for the nitrogen (N23), the sulfur (S24), its associated oxygens (O25, O26), and the methyl carbon (C27), indicating a critical inability to recognize this common and important functional group. Overall, the model only succeeds in identifying some atoms within unsubstituted aromatic regions.

These validation cases reveal fundamental limitations that highlight critical areas for future development. The consistent failure to recognize flexible aliphatic regions, essential functional groups like sulfonamides, and the bias toward aromatic fragment predictions suggest that the current voxelization approach may inadequately capture the diverse chemical environments encountered in practical drug optimization. Future work should prioritize enhancing training datasets with better representation of saturated and flexible molecular regions.

5.3 Future work and perspectives.

The systematic evaluation of our CNN-based fragment classification methodology has revealed both promising capabilities and fundamental limitations that define clear trajectories for future research. While achieving 67% top-1 accuracy and demonstrating robust generalization to novel protein families, the validation studies on clinically relevant targets exposed critical gaps that must be addressed to realize the full potential of classification-based approaches in structure-based drug design. The most pressing limitation revealed through our validation cases (HCV NS5B polymerase inhibitor) is the model's systematic failure to recognize flexible aliphatic regions and complex functional groups. Future work should prioritize developing augmented training datasets that explicitly balance aromatic and aliphatic fragment representation. This could be achieved through targeted mining of PDB structures containing

saturated heterocycles, flexible linkers, and sp^3-rich molecules, potentially incorporating structures from fragment screening campaigns where such moieties are overrepresented.

While the current voxelization approach and Gaussian smoothing showed good performance, incorporating explicit pharmacophoric features like hydrogen bond donors/acceptors, hydrophobic centers, and atomic charges as additional voxel channels could provide the model with more chemically relevant information for distinguishing similar functional groups. In addition, the difficulty in distinguishing chemically similar fragments within merged classes suggests the need for hierarchical learning approaches. Future architectures could implement a two-stage classification system: first identifying broad chemical families (aromatic systems, aliphatic chains, heteroatom-rich groups), then performing fine-grained classification within each family. This hierarchical approach could be coupled with graph neural network components to explicitly model covalent connectivity patterns. Finally, to address the practical challenge of applying this model to empty pockets where no parent ligand serves to center the voxel grid, future work will focus on developing a systematic scanning algorithm. By sampling multiple grid centers across the accessible binding pocket volume, the model can construct a comprehensive 3D prediction map, allowing it to identify specific sub-pockets and optimal fragment positions independent of prior ligand information. While our classification framework ensures chemical validity, combining it with controlled generative models could expand the accessible chemical space while maintaining interpretability. Future work should explore hybrid architectures where the classification model provides fragment anchors, and conditional generative models elaborate these fragments into complete molecules.

Beyond architectural improvements, translating this classification framework into practical drug discovery workflows requires addressing the critical gap between computational predictions and synthetic feasibility. A fundamental limitation of current deep learning approaches in SBDD is their tendency to function as "black boxes" that generate predictions without meaningful chemist interaction, thereby undermining the chemical intuition essential for rational drug design. To address this challenge, we are developing an interactive graphical user interface integrated with PyMOL that positions computational predictions as collaborative tools rather than autonomous decision-makers. This interface will enable medicinal chemists to select specific binding site

regions for fragment prediction, visualize the probability distribution across top predictions. By presenting predictions with associated confidence scores and allowing users to guide the model toward synthetically accessible and chemically logical modifications, this approach preserves the interpretability advantages of classification-based methods while allowing chemists to leverage their domain expertise in fragment selection and molecular elaboration strategies.

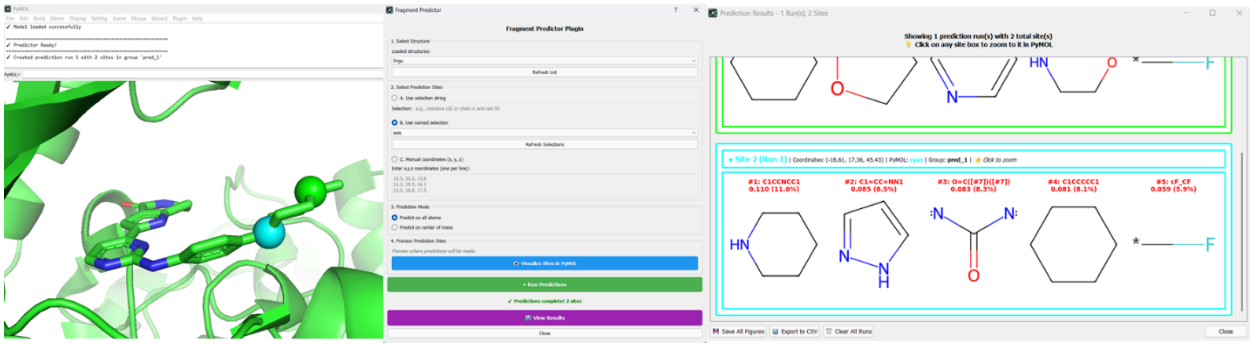


Figure 44. Interactive PyMOL-integrated graphical user interface for fragment prediction and visualization. The interface comprises three components: (Left) PyMOL visualization displaying the protein binding site (green cartoon) with reference ligand (blue sticks) and user-selected prediction point (cyan and green spheres). (Center) Control panel for model selection, prediction site specification, and execution controls. (Right) Results panel showing the molecular context (top) and ranked fragment predictions (bottom) with 2D structures, probability scores, and confidence percentages.

5.4 Conclusions

We presented a classification-based methodology for fragment prediction in structure-based drug design, addressing fundamental challenges in current deep learning approaches through systematic integration of chemical knowledge, protein diversity considerations, and rigorous optimization strategies. Through the development and validation of our CNN-based framework, we have demonstrated both the potential and current limitations of classification approaches for practical drug discovery applications.

The integration of Murcko scaffold decomposition with Ertl functional group analysis ensures comprehensive coverage of chemical space, capturing both rigid ring systems and flexible

heteroatom-containing moieties that single-method approaches miss. This dual strategy, coupled with our MAX10K balancing approach that maximizes protein diversity while addressing class imbalance, establishes a robust framework for fragment-based machine learning that others can build upon. The systematic hyperparameter optimization revealed critical insights about model behavior and dataset composition. The implementation of Gaussian smoothing (r=1.75 Å) for voxel representations proved essential for achieving reasonable performance with receptor-only inputs. The model's ability to maintain consistent performance (64% accuracy) under sequence-based splitting validates its capacity to generalize to novel protein families, a critical requirement for practical deployment. This robustness, combined with 80% top-5 accuracy, suggests that the learned representations capture meaningful protein-fragment interaction patterns that transcend specific protein families. The successful identification of conserved pharmacophoric elements in SARS-CoV-2 3CL protease and ERK2 kinase inhibitors further demonstrates the model's utility for recognizing established drug-like fragments.

However, our validation studies have also exposed fundamental limitations that constrain current applicability. The systematic failure to recognize flexible aliphatic regions, complex functional groups like sulfonamides, and conformationally diverse linkers reveals that the fragment selection could benefit of a wider chemical space. Based on the benchmark on models with 60 classes the performance at top 1 and top-5 predictions are still with reasonable performance, allowing us to include more classes and cover more chemical space.

The implications of this research extend beyond the specific models developed. The dataset curation pipeline, combining SASA filtering with protein sequence clustering, provides a template for creating unbiased training sets from the structurally biased PDB. The chemistry-informed fragment merging strategy demonstrates how domain expertise can guide machine learning architecture decisions, bridging the gap between computational methods and medicinal chemistry intuition.

5.5. Computational details

5.5.1. Protein-Ligand Complex Dataset: PDB Mining and Quality Filtering Criteria

Our methodology began with the construction of a comprehensive dataset, which involved systematically mining the Protein Data Bank (PDB) for high-quality protein-ligand complexes suitable for structural analysis. To ensure the reliability of atomic coordinates and the overall structural integrity, an initial quality filter was applied to retain only X-ray crystallographic structures with a resolution better than 2.5 Å. The scope of the dataset was precisely defined to include only protein-small molecule and protein-DNA-small molecule complexes, thereby excluding structures that consist purely of proteins or nucleic acids. Within these selected complexes, all small molecule entries were initially extracted as potential ligands, a category that included cocrystallized molecules. However, this initial pool was refined by subsequently removing ions and any molecules that failed the kekulization process in RDKit, a step necessary to ensure valid chemical representations. Finally, the canonical SMILES strings for the ligands and the FASTA sequences for the corresponding proteins were programmatically extracted using the RCSB PDB API.

5.5.2. Dual Molecular Fragmentation Strategy for Comprehensive Chemical Space Coverage

To decompose ligands into chemically meaningful fragments suitable for machine learning classification, we implemented two complementary fragmentation algorithms that capture distinct yet overlapping aspects of molecular structure, thereby ensuring comprehensive coverage of the chemical space relevant to structure-based drug design. This dual-strategy approach was specifically designed to address the limitations inherent in single-method fragmentation, where relying exclusively on either scaffold-based or functional group-based decomposition could result in incomplete representation of molecular diversity.

5.5.3. Murcko Scaffold Analysis for Ring Systems and Core Structures

The first algorithm employed was the Murcko scaffold decomposition, implemented using RDKit's Scaffold Network tool. By applying the default initialization parameters, we generated

explicit Bemis-Murcko scaffolds , a process that systematically prunes terminal side chains to identify and extract the core ring systems and their connecting linkers from each ligand. Crucially, this method preserves the original atom identities, hybridization states, and bond orders, rather than reducing the structure to a generic carbon framework. A key advantage of this approach is its ability to preserve cyclic and aromatic systems as intact structural units. The algorithm effectively generates a hierarchy of scaffolds, ranging from simple monocyclic rings like benzene and pyridine to complex, multi-ring fused systems, all while maintaining their fundamental core connectivity patterns.

5.5.4.Ertl Algorithm for Functional Group Decomposition

In parallel, we applied the Ertl fragmentation algorithm to identify functional groups based on heteroatom-containing moieties. This rule-based method systematically severs specific non-ring bonds to isolate key chemical groups like amines, ethers, and carbonyls. According to its definition, the algorithm's rules specifically target the cleavage of (1) single bonds between a carbon and a heteroatom and (2) single bonds connecting a carbonyl group to an adjacent alpha-carbon. This process is designed to break down molecules into smaller, chemically intuitive pieces while preserving the core identity of the functional groups, providing a complementary view to the rigid scaffolds identified by the Murcko analysis.

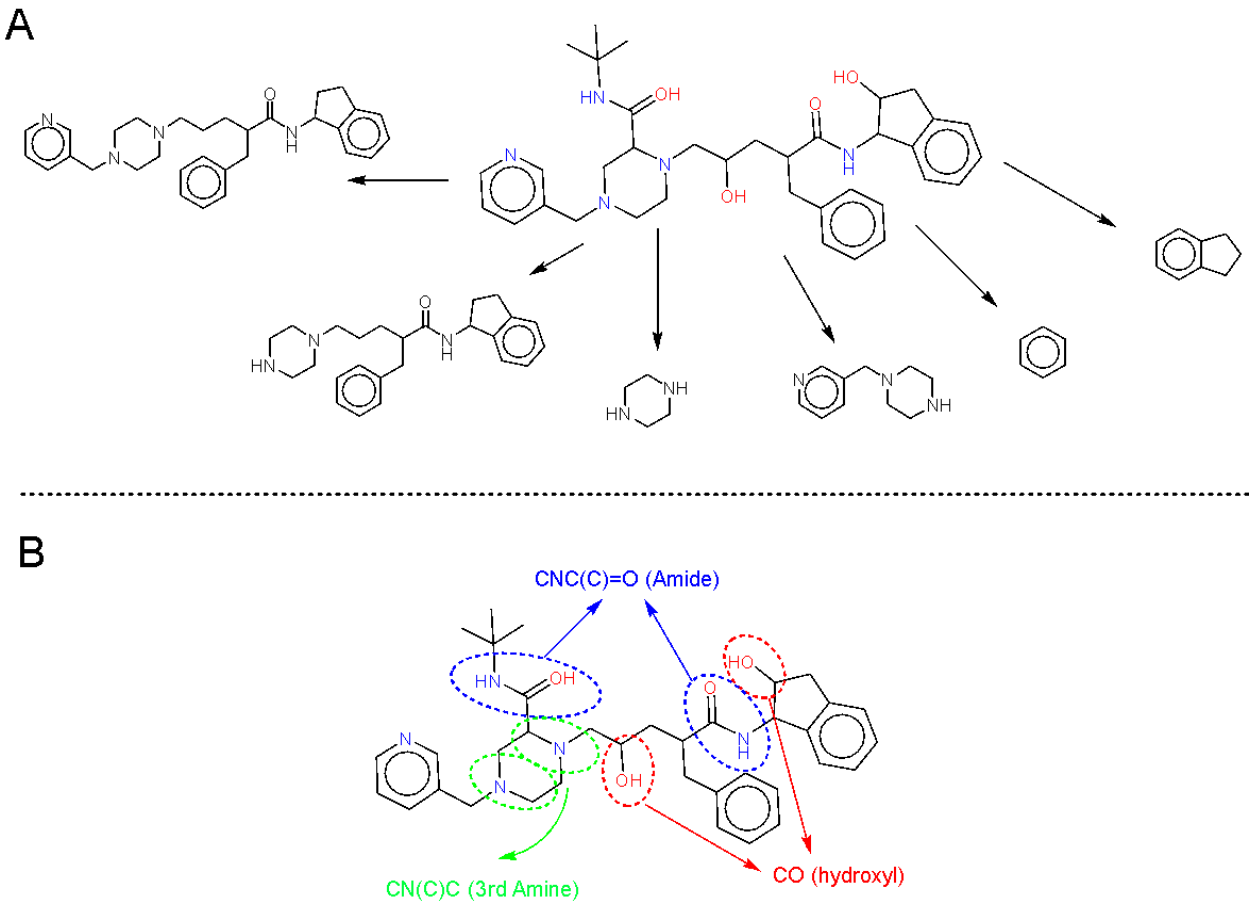


Figure 45. Dual molecular fragmentation strategy for comprehensive chemical space coverage. (A) Murcko scaffold decomposition systematically extracts rigid ring systems and aromatic cores as intact structural units from a representative PDB ligand. (B) Ertl fragmentation identifies heteroatom-containing functional groups through systematic bond cleavage, targeting amide groups [CNC(C)=O], tertiary amines [CN(C)C], and hydroxyl functionalities [CO]. The complementary approaches ensure comprehensive coverage of chemical space by capturing distinct aspects of molecular structure through scaffold-based and functional group-based decomposition.

5.5.5.Redundancy Reduction Through Chemical and Protein Sequence Clustering

To create a well-defined and non-redundant dataset for training, we applied a series of filtering and clustering steps to both the chemical fragments and the protein targets. These procedures were designed to normalize fragment representations, reduce bias from overrepresented protein families, and ensure that only fragments from well-defined binding pockets were included.

5.5.6. Fragment Chemical Similarity Clustering and Equivalence Mapping

Fragments generated by both algorithms were subjected to chemical similarity analysis to identify redundant representations. SMILES strings representing identical chemical groups with minor variations (e.g., CO and C[O-] for hydroxyl groups) were mapped to equivalent classes. We performed hierarchical clustering using Tanimoto similarity with a threshold 1 to group chemically equivalent fragments. Additionally, we curated a specialized list of medicinal chemistry-relevant fragments not captured by automated fragmentation, ensuring curated coverage of common modification sites used in drug optimization.

5.5.7. Protein Sequence Identity Clustering at 50% Threshold

To prevent model bias towards heavily studied protein families and improve generalization, we clustered all protein sequences using the MMseqs2 software. For proteins to be grouped into the same cluster, they had to satisfy two key criteria. The primary condition was a minimum sequence identity of 50%. In addition, we enforced a strict sequence coverage threshold, which requires the alignment between two proteins to span at least 80% of their sequence length. This second constraint is critical as it ensures that proteins are clustered based on their overall structural similarity, preventing misleading groupings based on only short, conserved local domains.

5.5.8. Filtering of buried fragments with Solvent-Accessible Surface Area

We implemented a physicochemical filter to select for fragments located in buried binding pockets and exclude those on exposed protein surfaces. For each complex, we calculated the Solvent-Accessible Surface Area (SASA) for the fragment's atoms in both the protein-bound state and an unbound (isolated) state, using a 1.4 Å probe radius. The relative SASA was then computed as the ratio of the bound SASA to the unbound SASA. Only complexes with a relative SASA below 0.25 were kept for the final dataset, as this threshold effectively selects for fragments deeply situated within a binding cavity. The SASA values were calculated using the Shrake & Rupley algorithm implemented in the Biopython package.

5.5.9. Three-Dimensional Voxel Representation of Protein-Fragment Binding Sites

For each protein-fragment complex, we generated a voxelized, or grid-based, representation of the local atomic environment to serve as the input for the 3D convolutional neural network.The process began by defining a three-dimensional cubic grid, typically 12×12×12 Å³, centered on the fragment's center of mass. Each grid was structured as a multi-channel tensor to encode the presence and type of atoms within its boundaries. Specifically, we used six distinct channels: three for the protein receptor atoms (carbon, nitrogen, and oxygen) and three for the atoms of the parent ligand that were not part of the fragment itself (carbon, nitrogen, and oxygen). The fragment's own atoms were excluded from the input representation. The value assigned to each voxel was determined using one of two methods. The first was a simple binary indicator. The second method, implemented to account for the dynamic nature of the structures, used a smoothed density function. In this approach, each atom α provides a contribution cα to the value of its surrounding voxels, where the contribution diminishes with the distance d from the atom to the center of each voxel. The contribution is calculated as follows, using a cutoff distance of r=1.75A˚:

$$C_{\alpha} = \{1 \quad if\ d_i < 0.5\ A\ 0 \qquad if\ d_i > 4\ e^{-\frac{2d^2}{r^2}} \quad otherwise$$

The final value in a given voxel for any channel is the sum of these contributions from all relevant atoms. Finally, to make the model robust to orientation, we applied a data augmentation step where each voxel sample was transformed through all 24 symmetric rotations of the cube, expanding the dataset and reducing orientation bias.

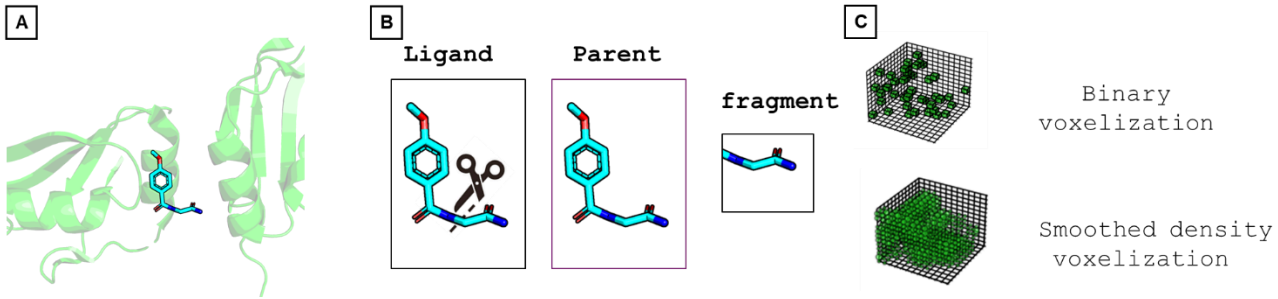


Figure 46. (A) Protein-ligand complex showing the local binding environment around a fragment of interest (highlighted in blue). (B) Decomposition of the complex into its constituent parts: the complete ligand, the parent molecule (ligand excluding the target fragment), and the isolated fragment used for classification. (C) Voxelization approaches showing binary representation (top) where each voxel contains discrete atomic presence indicators, and smoothed density representation (bottom) using a Gaussian function with r=1.75 Å

5.5.10. CNN Architecture Design: Fire Module-Based Classification Network

We implemented and compared three distinct 3D convolutional neural network (CNN) architectures to evaluate the impact of complexity on fragment classification. A detailed breakdown of each architecture and the total number of trainable parameters for each model can be found in the Supporting Information (Tables S2-S3). CNNScore, is based on the architecture from the KDEEP model, which adapts the highly efficient SqueezeNet for 3D structural data. This design employs signature "Fire Modules"—consisting of a 1x1 "squeeze" convolutional layer that feeds into a mix of 1x1 and 3x3 "expand" layers—to capture rich, multi-scale spatial patterns from the 12×12×12×6 voxel inputs while remaining computationally efficient. The network uses Rectified Linear Units (ReLUs) as activation functions, followed by global average pooling and fully connected layers to produce the final probability distribution over the fragment classes.

To benchmark this complex design, we implemented two simpler variants using standard convolutional blocks. The first, CNNBase, consists of three sequential 3D Convolutional layer followed by Batch Normalization layers, and a final max pooling and fully connected. The second, CNNBase_1C, is a minimal model with only a single convolutional layer. Despite its structural simplicity, its large final linear layer gives it a substantially higher parameter count than the other two models. Comparing the performance of these three architectures allowed us to assess the importance of specialized modules versus standard convolutions for this classification task.

5.5.11. Training and Evaluation Strategy: Hyperparameter optimization and dataset splits

The final dataset was partitioned into training, validation, and test sets to facilitate model development and rigorous evaluation. We employed two distinct methodologies for creating

these splits: a random split, where samples were uniformly distributed, and a more stringent sequence-based split, where all complexes from the same protein sequence cluster were assigned exclusively to a single partition to prevent data leakage from homologous proteins. A key aspect of our methodology was also exploring dataset composition by testing three balancing strategies: an Unbalanced strategy using all available data; a Balanced approach that created a smaller dataset by retaining only high-frequency classes; and a hybrid Unbalanced MAX10K strategy that maximized protein diversity before oversampling classes to a cap of 10,000 samples. Using these various dataset configurations, we performed a systematic hyperparameter optimization guided by model performance on the validation set.

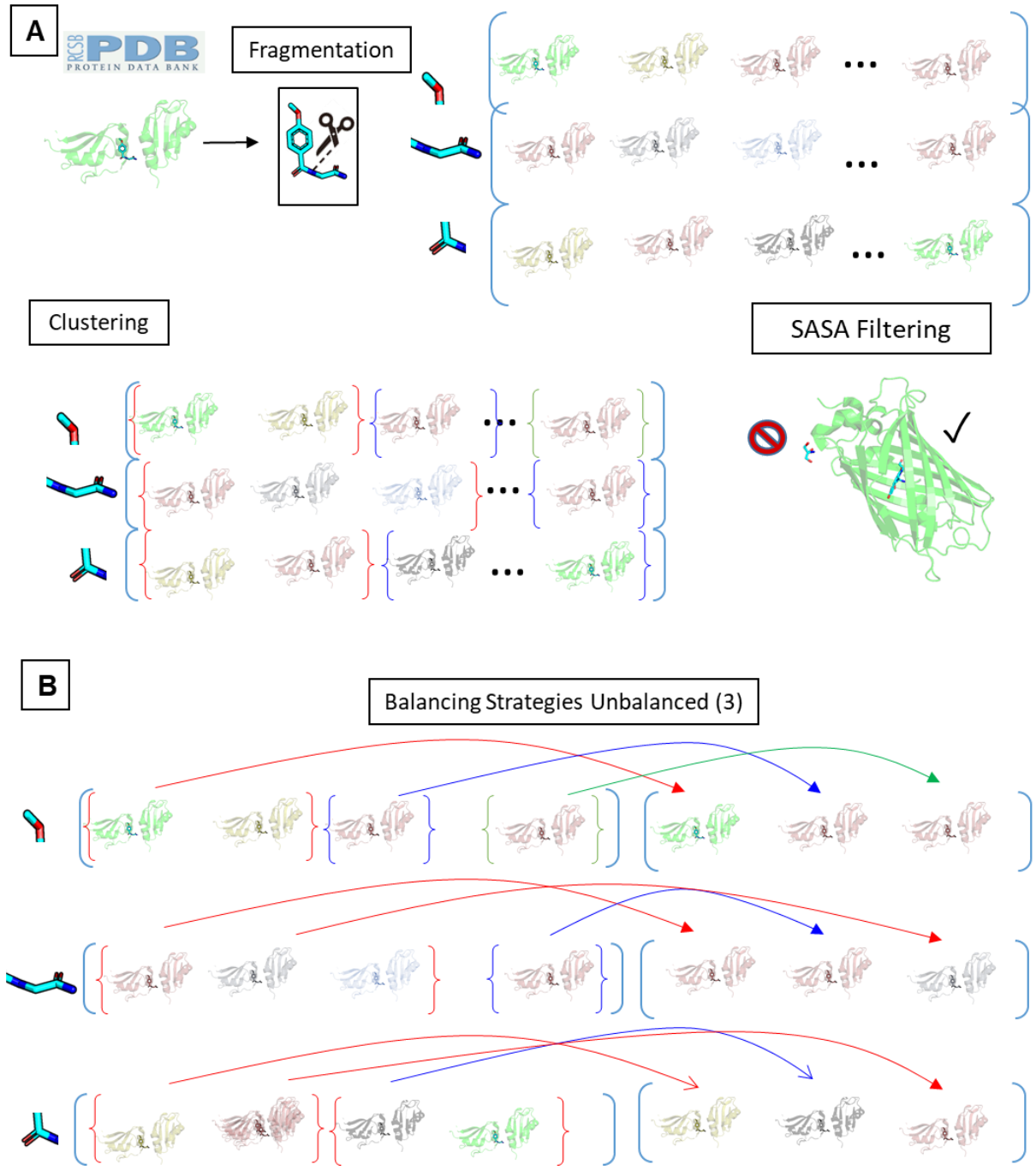


Figure 47. Comprehensive data preprocessing pipeline and MAX10K balancing strategy for fragment classification. (A) Sequential filtering workflow starting from PDB mining and fragmentation to generate protein-fragment pairs, followed by protein sequence clustering at 50% identity to group homologous structures, and SASA-based filtering (threshold <0.25) to retain only fragments occupying buried binding pockets while excluding surface-exposed interactions. (B) MAX10K balancing strategy implementation showing how protein clusters (colored brackets) for each fragment class are systematically sampled to achieve balanced datasets. The approach first prioritizes selecting one representative from each unique protein

cluster to maximize structural diversity, then supplements underrepresented classes by additional sampling from already-used clusters until reaching the 10,000 sample cap, ensuring both class balance and maximum protein diversity across the training dataset.

Then we procced to perform the hyperparameters optimization which includes the application of a Gaussian smoothing radius (r=1.75), the use of different balancing strategies and the optimization of paramters related to voxel representation, such as grid size (6 Å, 10 Å, and 12 Å) and; training parameters like the learning rate (tested from 1×10^{-5} to 1×10^{-3}) and batch size (32 and 64); and regularization techniques, such as label smoothing with factors of 0.0, 0.05, and 0.1. The models were trained by minimizing a categorical cross-entropy loss function, and their final performance was evaluated on the unseen test set. For this evaluation, we used a comprehensive set of metrics: top-1 accuracy was the primary metric for optimization, supplemented by top-K accuracy (with a focus on K=5) to assess practical utility. To understand performance at a granular level, per-class accuracy and the F1-score were calculated for individual fragments, and prediction entropy was used to quantify model confidence.

6. Conclusions

Through three complementary investigations spanning molecular dynamics simulations, alchemical free energy calculations, and deep learning approaches, we have addressed fundamental questions about nucleotide selection, discrimination mechanisms, and fragment-based drug design. These studies collectively advance our understanding of how polymerases achieve their remarkable fidelity while also revealing both the capabilities and limitations of current computational methodologies.

Our investigation of DNA polymerase β revealed a previously underappreciated mechanism for nucleotide discrimination operating through distal residues. The R283A mutation, despite its substantial distance from the metal centers, reduces fidelity by nearly two orders of magnitude. Through extensive molecular dynamics simulations, we demonstrated that R283 forms critical hydrogen bonds with the N2 atom of the template guanine base, stabilizing the reactive conformation required for catalysis.

The loss of this interaction in the R283A mutant creates additional volume within the binding site, causing Y271 to adopt alternative conformations that eliminate its ability to distinguish between matched and mismatched base pairs. This conformational reorganization produces structurally similar binding sites for both correct (dG:dCTP) and incorrect (dG:dATP) pairs, explaining the dramatic loss of selectivity observed experimentally. Importantly, structural alignment across polymerase X-family members from diverse organisms revealed that this arginine residue and its spatial relationship to the nascent base pair are highly conserved, suggesting a universal mechanism applicable to DNA polymerases λ, μ, and TdT.

Our alchemical free energy calculations successfully reproduced experimental binding affinity trends, achieving relative binding free energies within 2 kcal/mol of experimental values for wild-type systems. This validation demonstrates that relative binding free energy calculations can serve as valuable tools for understanding mutation effects in polymerase-nucleotide complexes, complementing kinetic and structural experiments.

The ribonucleotide discrimination study exposed critical limitations in current molecular mechanics force fields while establishing benchmarks for future development. Our systematic

comparison of GAFF2 and SHAW force fields using the Alchemical Transfer Method revealed that general-purpose parametrizations fail for the most relevant mutants, Y271A and Y416A in Pol β and Pol RB69, respectively. Those simulations erroneously suggested that these DNA polymerases prefer ribonucleotides over deoxyribonucleotides. The specialized SHAW force field, developed through extensive quantum mechanical calculations on nucleotide systems, improved overall performance substantially. SHAW correctly predicted the qualitative preference for dCTP over rCTP in both wild-type and mutant DNA polymerases, demonstrating that force field accuracy directly impacts the reliability of free energy predictions. This success validates the investment in nucleotide-specific parametrization and provides confidence in applying these force fields to wild-type systems and conservative mutations.

These failures motivated our proposal to develop Neural Network Potential/Molecular Mechanics (NNP/MM) hybrid methods specifically for nucleotide systems. We initiated this effort by curating a comprehensive training dataset of 824 nucleotide molecules spanning diverse phosphorylation states, sugar modifications, and nucleobase variants. This dataset systematically samples the conformational and chemical space relevant to polymerase-nucleotide interactions while properly accounting for the high charge density characteristic of triphosphate groups. The anticipated NNP/MM approach should achieve quantum mechanical accuracy for the nucleotide and its immediate environment while maintaining computational efficiency through classical treatment of bulk protein and solvent.

Our classification-based approach to fragment prediction established a practical framework that balances accuracy, interpretability, and chemical validity. The final model achieved 67% top-1 accuracy and 80% top-5 accuracy across 30 chemically diverse fragment classes, performance levels that compare favorably with existing fragment-based methods while providing superior interpretability through explicit fragment identification rather than fingerprint-based predictions.

Three methodological innovations proved essential for achieving robust generalization. First, our dual fragmentation strategy combining Murcko scaffold decomposition with Ertl functional group analysis ensured comprehensive coverage of chemical space, capturing both rigid aromatic systems and flexible heteroatom-containing moieties that single-method approaches miss. Second, the MAX10K balancing strategy maximized protein diversity by prioritizing sampling across different sequence clusters before resampling within clusters, effectively addressing the

inherent bias in the Protein Data Bank toward heavily studied protein families. Third, chemistry-informed fragment merging consolidated equivalent functional groups (such as aromatic versus aliphatic amines) while maintaining distinct classes for chemically different moieties, reducing unnecessary granularity without sacrificing discriminative power.

Validation on clinically relevant targets revealed both the method's strengths and its current limitations. For SARS-CoV-2 3CL protease inhibitors, the model successfully identified conserved pharmacophoric elements including the triazinedione core and trifluorobenzyl substituents. For ERK2 kinase inhibitors, predictions accurately captured the pyrimidine-pyrrololactam scaffold and morpholine substituent patterns that characterize this series. However, analysis of HCV NS5B polymerase inhibitors exposed systematic failures in recognizing flexible aliphatic linkers, sulfonamide groups, and tert-butyl substituents. Additionally, the model maintained consistent performance under sequence-based data splitting, achieving 64% accuracy when evaluated on completely novel protein families. This robustness validates that the learned representations capture genuine protein-fragment interaction patterns rather than memorizing family-specific features, a crucial requirement for practical deployment in drug discovery campaigns targeting understudied proteins.

In overall, this work advances polymerase research by demonstrating how computational methods can provide mechanistic insights inaccessible to experiment alone. The discovery that R283 influences fidelity through dynamic effects on active site geometry illustrates how mutations far from catalytic residues can dramatically impact function. The systematic force field benchmarking establishes standards for reliability in nucleotide simulations, benefiting researchers studying DNA repair, mutagenesis, and antiviral drug development. For drug discovery more broadly, this thesis contributes validated methodologies spanning multiple scales. Free energy calculations, when combined with appropriate force fields, can predict mutation effects and guide medicinal chemistry with reasonable confidence. Deep learning approaches can suggest fragments for lead optimization while maintaining chemical validity and interpretability—critical requirements for practical deployment. The emphasis on rigorous validation against experimental data and assessment of limitations provides a model for responsible integration of computational methods into discovery pipelines.

The path forward requires continued development of both approaches in tandem. Neural network potentials promise to combine quantum mechanical accuracy with molecular dynamics efficiency, potentially resolving long-standing force field limitations. Larger, better-curated datasets will enable more sophisticated machine learning architectures to learn genuinely transferable patterns. Integration of these methods into unified workflows will accelerate discovery timelines and explore function and inhibition more completely.

Polymerases remain central to life, disease, and biotechnology. Their investigation through computational methods has revealed fundamental principles of molecular recognition, catalysis, and drug design. As methods mature and computational power increases, we anticipate that the approaches combining molecular dynamics simulations, free energy calculations and deep learning, will contribute to new therapeutic interventions against viral and bacterial infections, improved treatments for cancers dependent on DNA repair pathways, and engineered polymerases with novel capabilities for synthetic biology.

7. Supplementary information

7.1. Chapter 3 Supplementary information

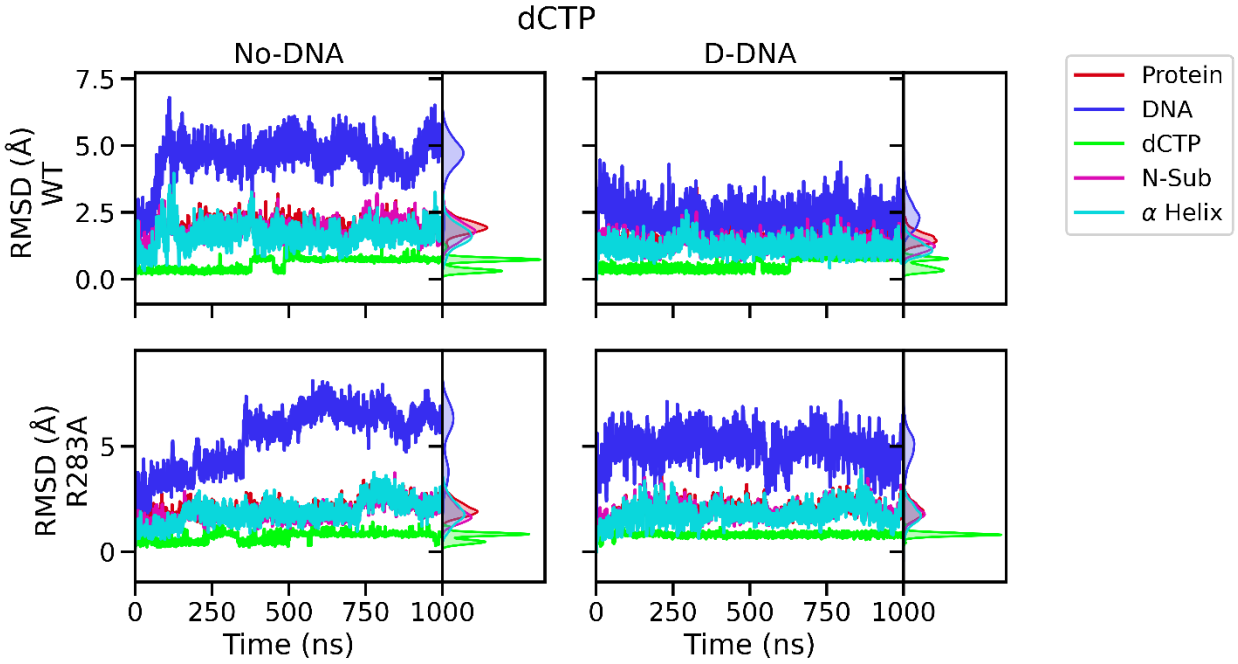


Figure S1 RMSD for Ca in protein backbone, DNA phosphate backbone , dNTP, N Subdomain and α Helix in N Subdomain for dG:dCTP-DNA-Pol β.

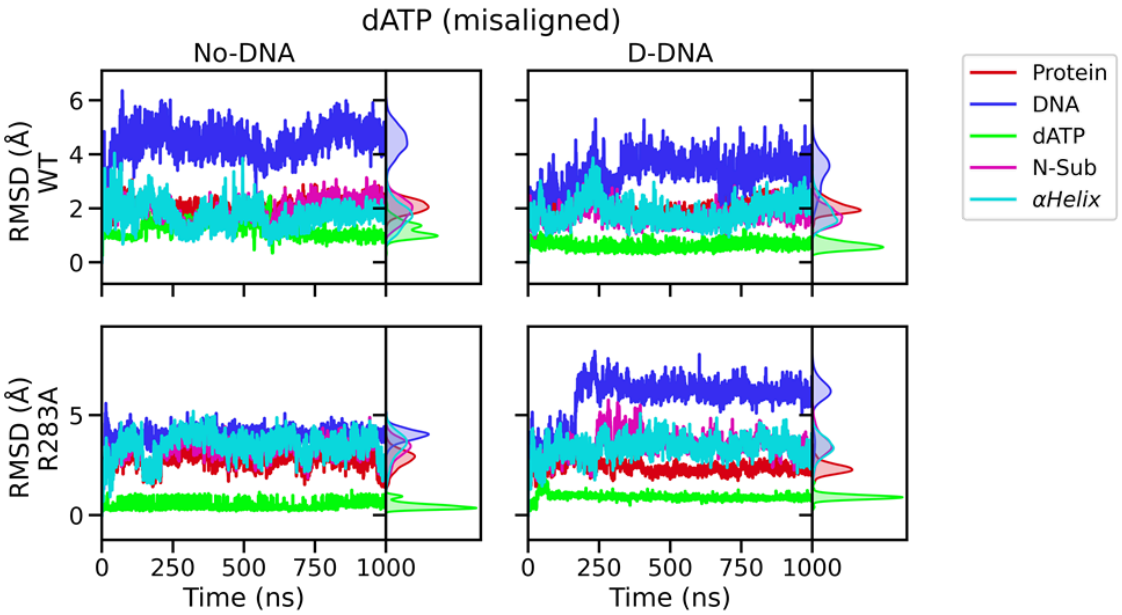


Figure S2. RMSD for Ca in protein backbone, DNA phosphate backbone , dNTP, N Subdomain and α Helix in N Subdomain for dG:dATP-DNA-Pol β with the catalytic site with misaligned moieties.

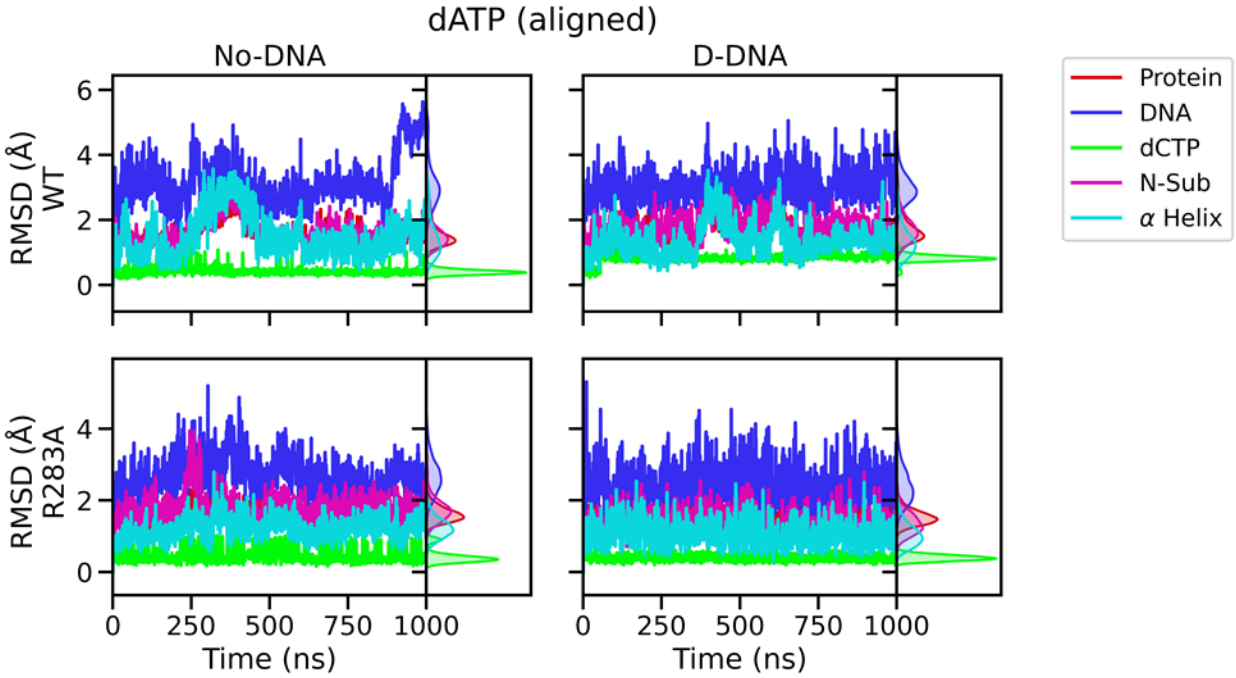


Figure S3 RMSD for Ca in protein backbone, DNA phosphate backbone , dNTP, N Subdomain and α Helix in N Subdomain for dG:dATP-DNA-Pol β with the catalytic site with aligned moieties.

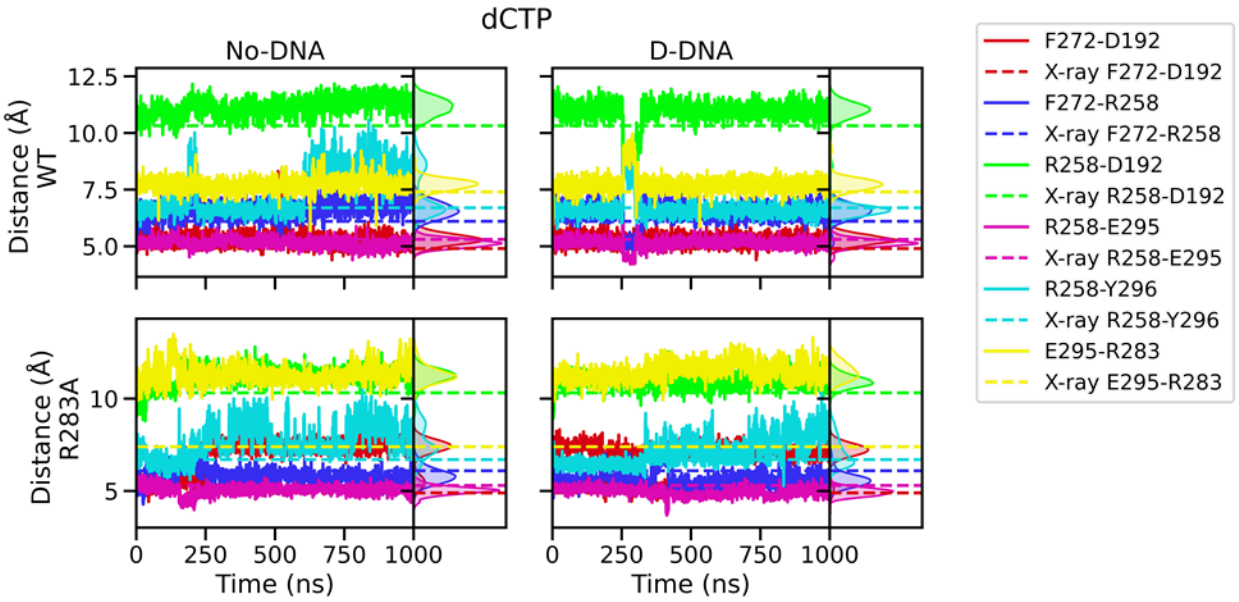


Figure S4. Relevant distances of Close conformation in the system dG:dCTP-DNA-Pol β (matched).

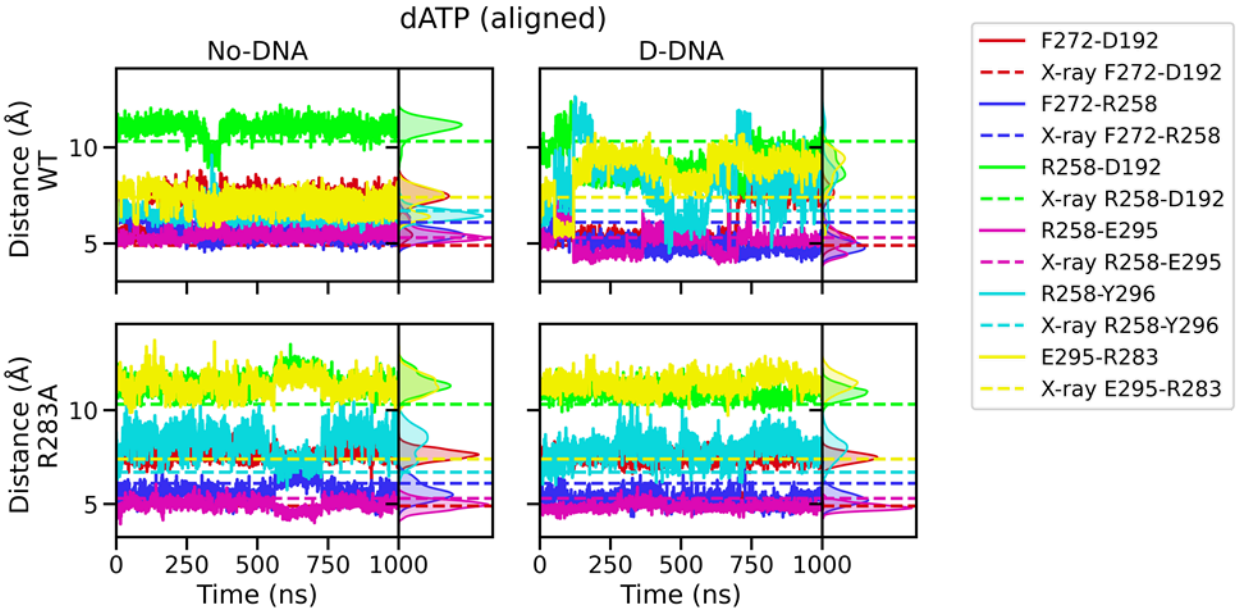


Figure S5. Relevant distances of Close conformation in the system dG:dATP-DNA-Pol β (mismatch with align catalytic center).

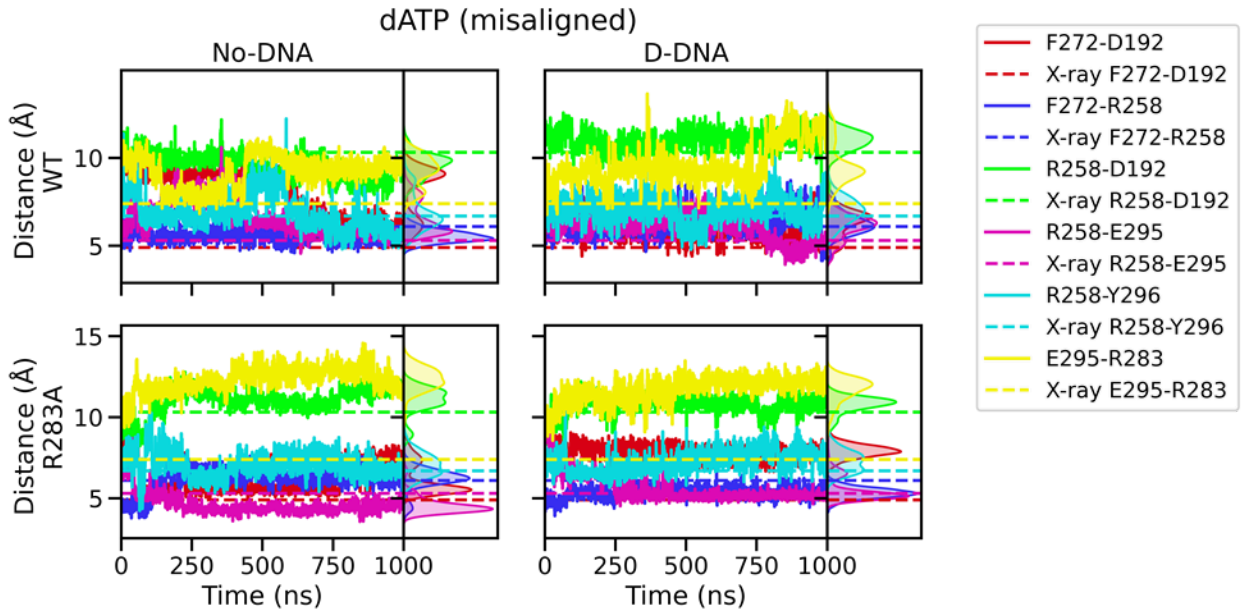


Figure S6. Relevant distances of Close conformation in the system dG:dATP-DNA-Pol β(mismatch with misalign catalytic center).

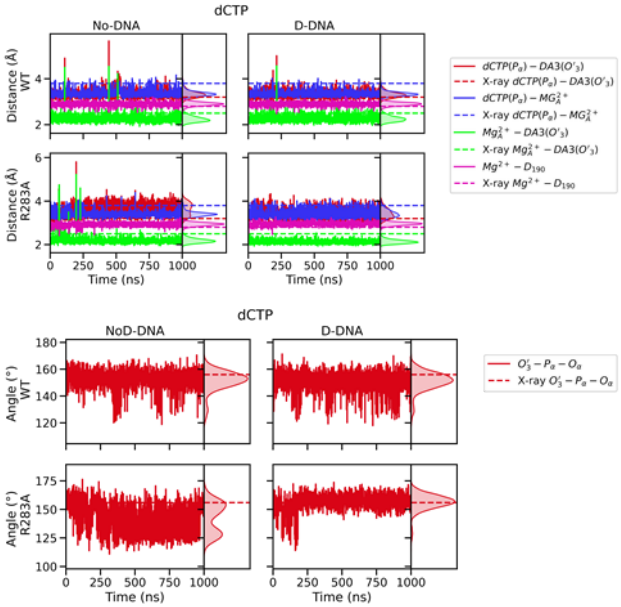


Figure S7. dG:dCTP-DNA-Pol β time series of geometrical parameters defining a aligned catalytic center. Right coordination with MgA to O3 and Pα Left, angle of nucleophilic attack.

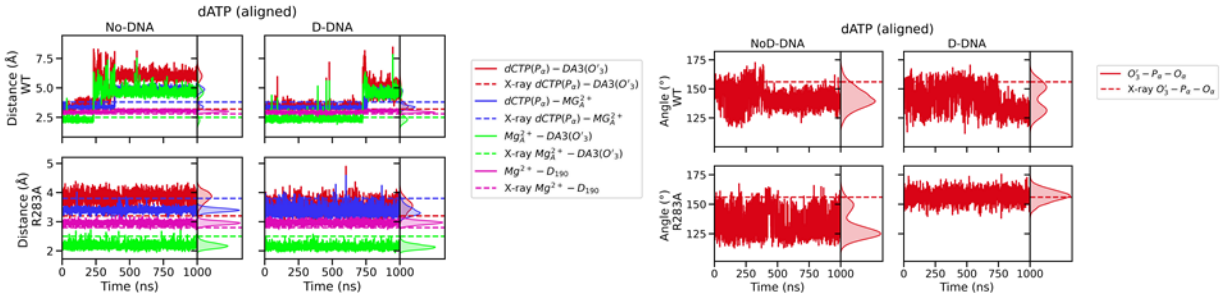


Figure S8. dG:dATP-DNA-Pol β time series of geometrical parameters for defining a aligned catalytic center. Right coordination with with MgA to O3 and Pα Left, angle of nucleophilic attack.

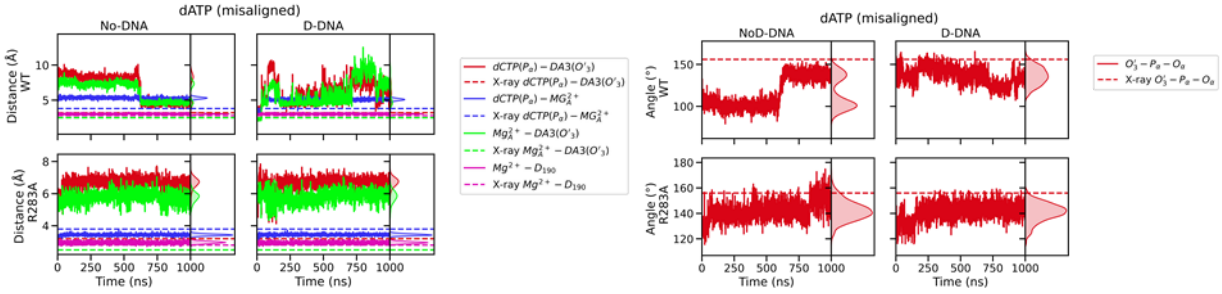


Figure S9. dG:dATP-DNA-Pol β time series of geometrical parameters for defining a aligned catalytic center. Right coordination with MgA to O3 and Pα Left, angle of nucleophilic attack.

Table S1. Hydrogen Bond occupancies for R283 in WT systems.

Interaction	dG:dCTP-WT (NoD-DNA)	dG:dCTP-WT (D-DNA)	dG:dATP-WT (NoD-DNA) (Aligned)	dG:dATP-WT (D-DNA) (Aligned)	dG:dATP-WT (NoD-DNA) (Misaligned)	dG:dATP-WT (D-DNA) (Misaligned)
ASN279@O ARG283@N	1.0	1.0	1.0	0.8	1.0	1.0
ARG283@NH1 DG(n)@N2	0.6	0.5	0.9	-	-	-
ARG283@NH2 DG(n)@N2	0.2	0.4	-	-	-	-
ASN279@OD1 ARG283@NH2	0.7	0.8	-	-	-	-
ARG283@O LEU287@N	0.8	0.7	0.4	0.7	0.6	0.9
DT(n-1)@O4' ARG283@NH1	0.5	0.5	-	-	-	-
ASN279@OD1 ARG283@NE	0.2	0.3	-	-	-	-
GLU295@OE2 ARG283@NH2	-	-	0.3	-	-	-
ARG283@NE DG(n)@N2	-	-	-	0.2	-	-
ASN294@O ARG283@NH2	-	-	-	0.3	-	-
ASN279@O ARG283@NH1	-	-	-	0.8	-	-
DG(n)@OP2 ARG283@NE	-	-	-	-	0.4	-
ASN279@OD1 ARG283@NH1	-	-	-	-	1.0	-
DG(n)@OP1 ARG283@NH2	-	-	-	-	0.5	-
DG(n)@OP1 ARG283@NE	-	-	-	-	0.5	-
DG(n)@O5' ARG283@NH2	-	-	-	-	0.5	-
DT(n-1)@OP1 ARG283@NH2	-	-	-	-	-	0.2
DG(n)@O3' ARG283@NH1	-	-	-	-	-	0.2

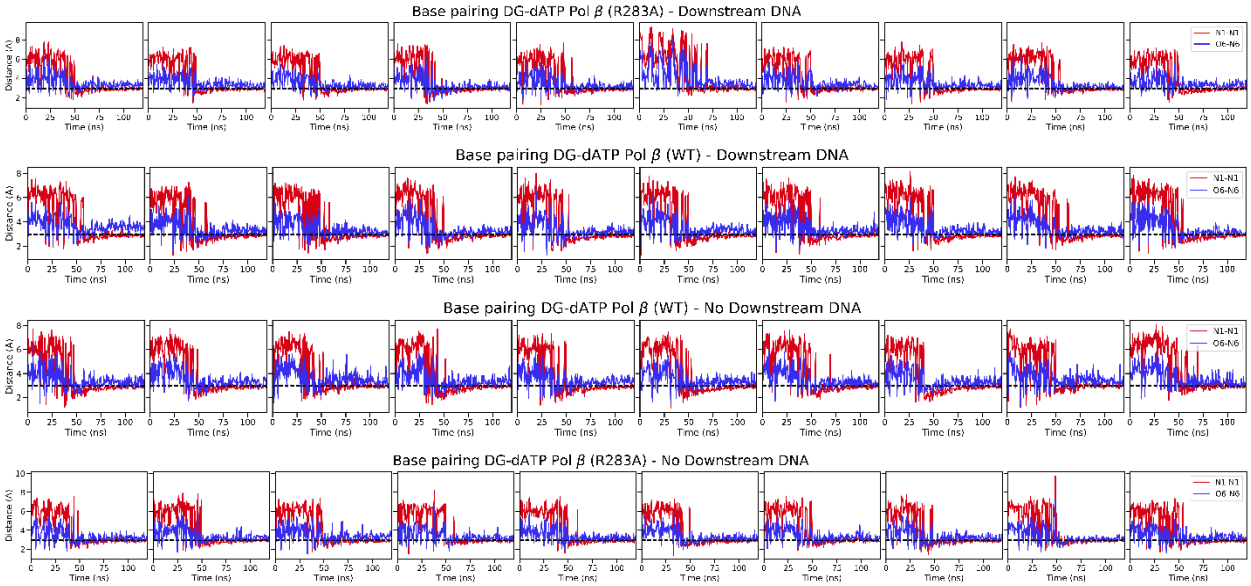


Figure S10 Hydrogen bonds for the dATP:dG during the alchemical free energy calculation. Dotted lines represent possible Hydrogen bond of mismatch based on the structure 4KLF.

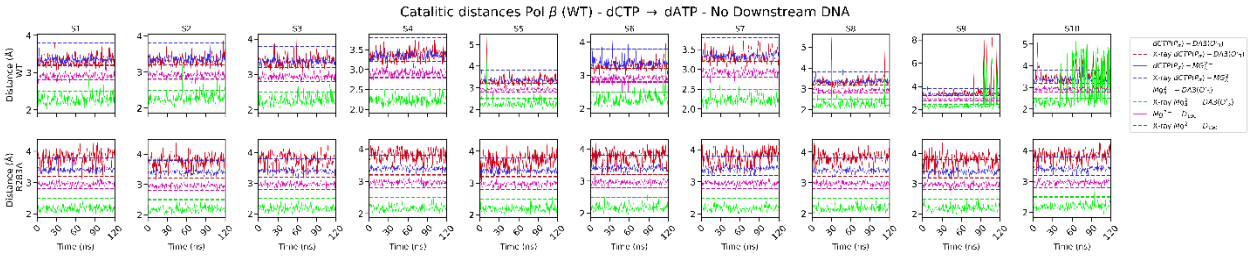


Figure S11. Catalytic distances for all replicas during the alchemical free energy calculation for the system Pol β–No Downstream

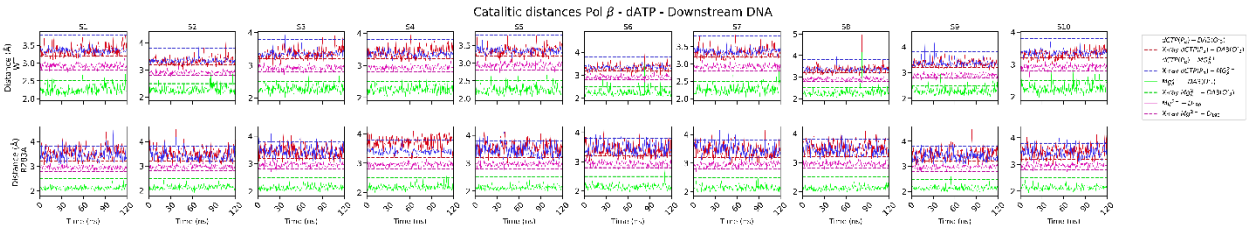


Figure S12. Catalytic distances for all replicas during the alchemical free energy calculation for the system Pol β–Downstream DNA

Table S2. Average ± s.d of relevant catalytic distances for all replicas in the alchemical free energy calculation.

System	Interaction	Ref	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
WT-D	dCTP-P DA3-O3	3.2	3.36 ± 0.15	3.33 ± 0.16	3.35 ± 0.15	3.40 ± 0.18	3.34 ± 0.13	3.32 ± 0.16	3.33 ± 0.14	3.35 ± 0.18	3.35 ± 0.14	3.31 ± 0.13
	dCTP-P MG_A2+	3.8	3.35 ± 0.07	3.34 ± 0.09	3.35 ± 0.08	3.36 ± 0.08	3.36 ± 0.07	3.35 ± 0.08	3.36 ± 0.08	3.35 ± 0.08	3.37 ± 0.14	3.35 ± 0.08
	MG-DA3-O3	2.5	2.27 ± 0.13	2.27 ± 0.15	2.27 ± 0.12	2.25 ± 0.13	2.25 ± 0.12	2.27 ± 0.21	2.26 ± 0.14	2.29 ± 0.18	2.29 ± 0.15	2.26 ± 0.15
	MG-D190	2.8	2.92 ± 0.07	2.91 ± 0.06	2.92 ± 0.07	2.93 ± 0.07	2.94 ± 0.06	2.92 ± 0.07	2.93 ± 0.06	2.92 ± 0.07	2.90 ± 0.07	2.93 ± 0.07
R283A-D	dCTP-P DA3-O3	3.2	3.56 ± 0.23	3.51 ± 0.24	3.45 ± 0.22	3.69 ± 0.22	3.55 ± 0.23	3.38 ± 0.22	3.45 ± 0.25	3.53 ± 0.22	3.49 ± 0.22	3.51 ± 0.23
	dCTP-P MG_A2+	3.8	3.46 ± 0.21	3.42 ± 0.21	3.45 ± 0.21	3.39 ± 0.09	3.46 ± 0.22	3.44 ± 0.20	3.47 ± 0.21	3.42 ± 0.17	3.40 ± 0.19	3.43 ± 0.19
	MG-DA3-O3	2.5	2.12 ± 0.08	2.13 ± 0.09	2.13 ± 0.09	2.18 ± 0.09	2.13 ± 0.08	2.15 ± 0.10	2.14 ± 0.09	2.12 ± 0.08	2.14 ± 0.09	2.13 ± 0.08
	MG-D190	2.8	2.95 ± 0.07	2.96 ± 0.07	2.96 ± 0.08	2.96 ± 0.07	2.97 ± 0.07	2.94 ± 0.08	2.95 ± 0.08	2.96 ± 0.07	2.98 ± 0.07	2.97 ± 0.08
WT-NoD	dCTP-P DA3-O3	3.2	3.33 ± 0.14	3.32 ± 0.14	3.37 ± 0.15	3.36 ± 0.15	3.34 ± 0.20	3.30 ± 0.13	3.33 ± 0.14	3.34 ± 0.23	3.62 ± 0.99	3.59 ± 0.39
	dCTP-P MG_A2+	3.8	3.35 ± 0.07	3.35 ± 0.09	3.36 ± 0.07	3.36 ± 0.07	3.35 ± 0.07	3.36 ± 0.12	3.35 ± 0.06	3.35 ± 0.10	3.36 ± 0.07	3.37 ± 0.10
	MG-DA3-O3	2.5	2.27 ± 0.13	2.28 ± 0.16	2.26 ± 0.14	2.24 ± 0.11	2.26 ± 0.19	2.27 ± 0.17	2.23 ± 0.11	2.28 ± 0.24	2.48 ± 0.87	2.89 ± 0.99
	MG-D190	2.8	2.92 ± 0.07	2.92 ± 0.07	2.93 ± 0.07	2.92 ± 0.07	2.92 ± 0.07	2.91 ± 0.08	2.93 ± 0.07	2.91 ± 0.08	2.92 ± 0.06	2.94 ± 0.07
R283A-NoD	dCTP-P DA3-O3	3.2	3.79 ± 0.22	3.77 ± 0.21	3.77 ± 0.22	3.81 ± 0.20	3.69 ± 0.23	3.77 ± 0.22	3.76 ± 0.22	3.83 ± 0.18	3.77 ± 0.23	3.84 ± 0.19
	dCTP-P MG_A2+	3.8	3.40 ± 0.08	3.40 ± 0.06	3.40 ± 0.06	3.40 ± 0.06	3.37 ± 0.07	3.40 ± 0.07	3.40 ± 0.07	3.40 ± 0.07	3.40 ± 0.08	3.40 ± 0.06
	MG-DA3-O3	2.5	2.17 ± 0.09	2.18 ± 0.09	2.17 ± 0.09	2.17 ± 0.09	2.19 ± 0.09	2.18 ± 0.11	2.18 ± 0.08	2.17 ± 0.10	2.20 ± 0.10	2.20 ± 0.10
	MG-D190	2.8	2.96 ± 0.06	2.96 ± 0.06	2.97 ± 0.06	2.96 ± 0.06	2.96 ± 0.07	2.97 ± 0.06	2.96 ± 0.06	2.98 ± 0.06	2.96 ± 0.07	2.97 ± 0.06

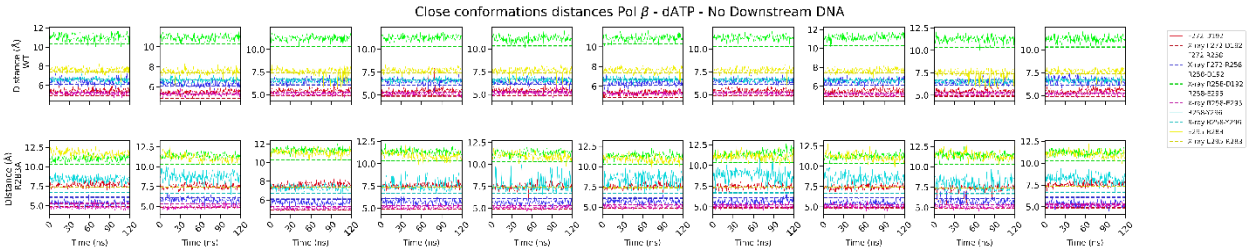


Figure S13. Relevant distances for close state for all replicas during the alchemical free energy calculation for the system Pol β– Downstream

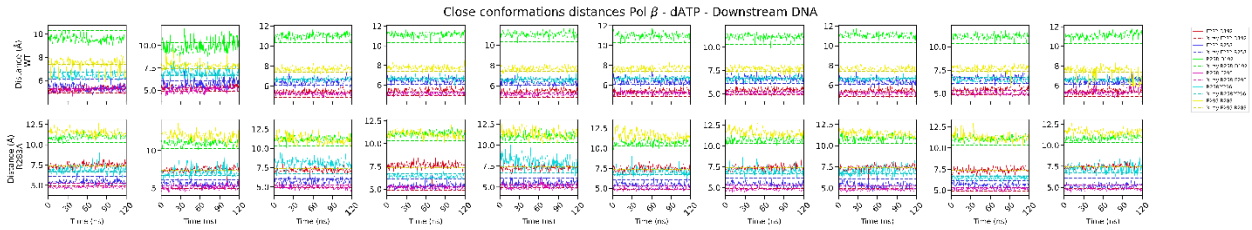


Figure S14. Relevant distances for close state for all replicas during the alchemical free energy calculation for the system Pol β–No Downstream

Table S 3. Average ± s.d for relevant distances for a close state in Pol β

System	Interaction	Ref	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
WT-D	F272-D192	4.9	5.24 ± 0.19	5.36 ± 0.23	5.45 ± 0.21	5.44 ± 0.23	5.44 ± 0.22	5.38 ± 0.21	5.41 ± 0.22	5.43 ± 0.21	5.40 ± 0.23	5.39 ± 0.22
	F272-R258	6.1	5.37 ± 0.28	5.71 ± 0.50	6.46 ± 0.27	6.46 ± 0.25	6.37 ± 0.24	6.45 ± 0.29	6.64 ± 0.28	6.39 ± 0.25	6.59 ± 0.25	6.42 ± 0.26
	R258-D192	10.32	9.58 ± 0.31	9.87 ± 0.78	11.03 ± 0.29	11.16 ± 0.26	11.07 ± 0.24	11.00 ± 0.27	11.13 ± 0.29	11.07 ± 0.24	11.10 ± 0.28	11.07 ± 0.29
	R258-E295	5.3	5.21 ± 0.21	5.06 ± 0.33	5.21 ± 0.15	5.09 ± 0.14	5.14 ± 0.16	5.23 ± 0.17	5.03 ± 0.15	5.14 ± 0.14	4.97 ± 0.15	5.36 ± 0.34
	R258-Y296	6.7	6.38 ± 0.28	7.03 ± 0.49	6.62 ± 0.19	6.58 ± 0.21	6.58 ± 0.21	6.64 ± 0.21	6.60 ± 0.24	6.62 ± 0.19	6.42 ± 0.21	6.60 ± 0.20
	E295-R283	7.4	7.50 ± 0.41	7.82 ± 0.35	7.73 ± 0.24	7.75 ± 0.22	7.66 ± 0.27	7.71 ± 0.22	7.71 ± 0.30	7.67 ± 0.19	7.67 ± 0.27	7.40 ± 0.64
R283A-D	F272-D192	4.9	7.46 ± 0.29	7.46 ± 0.33	7.17 ± 0.27	7.64 ± 0.35	7.28 ± 0.29	7.47 ± 0.37	7.37 ± 0.30	7.41 ± 0.33	7.32 ± 0.30	7.47 ± 0.26
	F272-R258	6.1	5.43 ± 0.30	5.47 ± 0.30	5.80 ± 0.34	5.36 ± 0.26	5.43 ± 0.33	5.44 ± 0.30	5.42 ± 0.26	5.36 ± 0.32	5.66 ± 0.24	5.21 ± 0.23
	R258-D192	10.32	10.96 ± 0.31	11.00 ± 0.34	11.22 ± 0.36	11.08 ± 0.25	11.03 ± 0.36	10.69 ± 0.32	10.73 ± 0.31	10.91 ± 0.32	11.07 ± 0.28	10.87 ± 0.28
	R258-E295	5.3	4.79 ± 0.16	5.03 ± 0.31	4.94 ± 0.21	5.19 ± 0.16	4.98 ± 0.27	4.90 ± 0.14	4.82 ± 0.16	4.81 ± 0.14	5.13 ± 0.15	4.80 ± 0.13
	R258-Y296	6.7	6.90 ± 0.45	7.12 ± 0.56	8.06 ± 0.54	6.41 ± 0.21	7.97 ± 0.72	7.01 ± 0.41	7.01 ± 0.42	6.90 ± 0.46	6.43 ± 0.23	7.22 ± 0.49
	E295-R283	7.4	11.45 ± 0.38	11.97 ± 0.52	11.14 ± 0.52	11.06 ± 0.36	11.37 ± 0.54	11.56 ± 0.52	11.55 ± 0.42	11.41 ± 0.35	11.26 ± 0.49	11.55 ± 0.35
WT-NoD	F272-D192	4.9	5.42 ± 0.22	5.47 ± 0.20	5.43 ± 0.21	5.47 ± 0.33	5.47 ± 0.22	5.38 ± 0.22	5.44 ± 0.20	5.48 ± 0.22	5.44 ± 0.22	5.52 ± 0.24
	F272-R258	6.1	6.45 ± 0.31	6.31 ± 0.26	6.53 ± 0.26	6.49 ± 0.27	6.40 ± 0.26	6.37 ± 0.31	6.49 ± 0.23	6.54 ± 0.27	6.47 ± 0.26	6.65 ± 0.33
	R258-D192	10.32	11.00 ± 0.33	10.93 ± 0.29	11.25 ± 0.30	11.26 ± 0.27	11.13 ± 0.32	10.90 ± 0.29	11.24 ± 0.27	11.17 ± 0.29	11.23 ± 0.35	11.18 ± 0.33
	R258-E295	5.3	5.16 ± 0.16	5.32 ± 0.17	5.07 ± 0.14	5.14 ± 0.16	5.15 ± 0.22	5.26 ± 0.17	5.07 ± 0.14	5.14 ± 0.15	5.11 ± 0.16	5.22 ± 0.18
	R258-Y296	6.7	6.58 ± 0.20	6.67 ± 0.21	6.53 ± 0.19	6.62 ± 0.21	6.57 ± 0.22	6.67 ± 0.20	6.55 ± 0.18	6.61 ± 0.17	6.50 ± 0.18	6.66 ± 0.21
	E295-R283	7.4	7.63 ± 0.27	7.52 ± 0.31	7.57 ± 0.42	7.71 ± 0.24	7.59 ± 0.38	7.61 ± 0.29	7.67 ± 0.26	7.67 ± 0.23	7.39 ± 0.58	7.66 ± 0.30
R283A-NoD	F272-D192	4.9	7.68 ± 0.27	7.53 ± 0.29	7.60 ± 0.28	7.54 ± 0.26	7.48 ± 0.29	7.56 ± 0.29	7.44 ± 0.27	7.46 ± 0.27	7.35 ± 0.40	7.59 ± 0.26
	F272-R258	6.1	5.35 ± 0.31	5.81 ± 0.31	5.71 ± 0.25	5.71 ± 0.26	5.68 ± 0.25	5.67 ± 0.31	5.63 ± 0.31	5.46 ± 0.29	5.74 ± 0.35	5.32 ± 0.29
	R258-D192	10.32	11.08 ± 0.32	11.39 ± 0.31	11.36 ± 0.26	11.34 ± 0.28	11.33 ± 0.29	11.36 ± 0.34	11.36 ± 0.37	11.24 ± 0.31	11.43 ± 0.34	11.21 ± 0.29
	R258-E295	5.3	4.81 ± 0.19	4.93 ± 0.20	5.09 ± 0.12	5.08 ± 0.18	5.06 ± 0.15	5.03 ± 0.19	5.08 ± 0.23	5.06 ± 0.21	5.04 ± 0.17	5.15 ± 0.19
	R258-Y296	6.7	8.42 ± 0.40	8.74 ± 0.59	7.30 ± 0.25	7.55 ± 0.55	7.59 ± 0.75	8.39 ± 0.71	8.84 ± 0.67	8.53 ± 0.61	7.85 ± 0.71	8.18 ± 0.52
	E295-R283	7.4	11.73 ± 0.39	11.14 ± 0.39	11.06 ± 0.31	10.96 ± 0.38	10.94 ± 0.38	10.91 ± 0.38	11.08 ± 0.45	11.10 ± 0.45	11.36 ± 0.43	11.01 ± 0.37

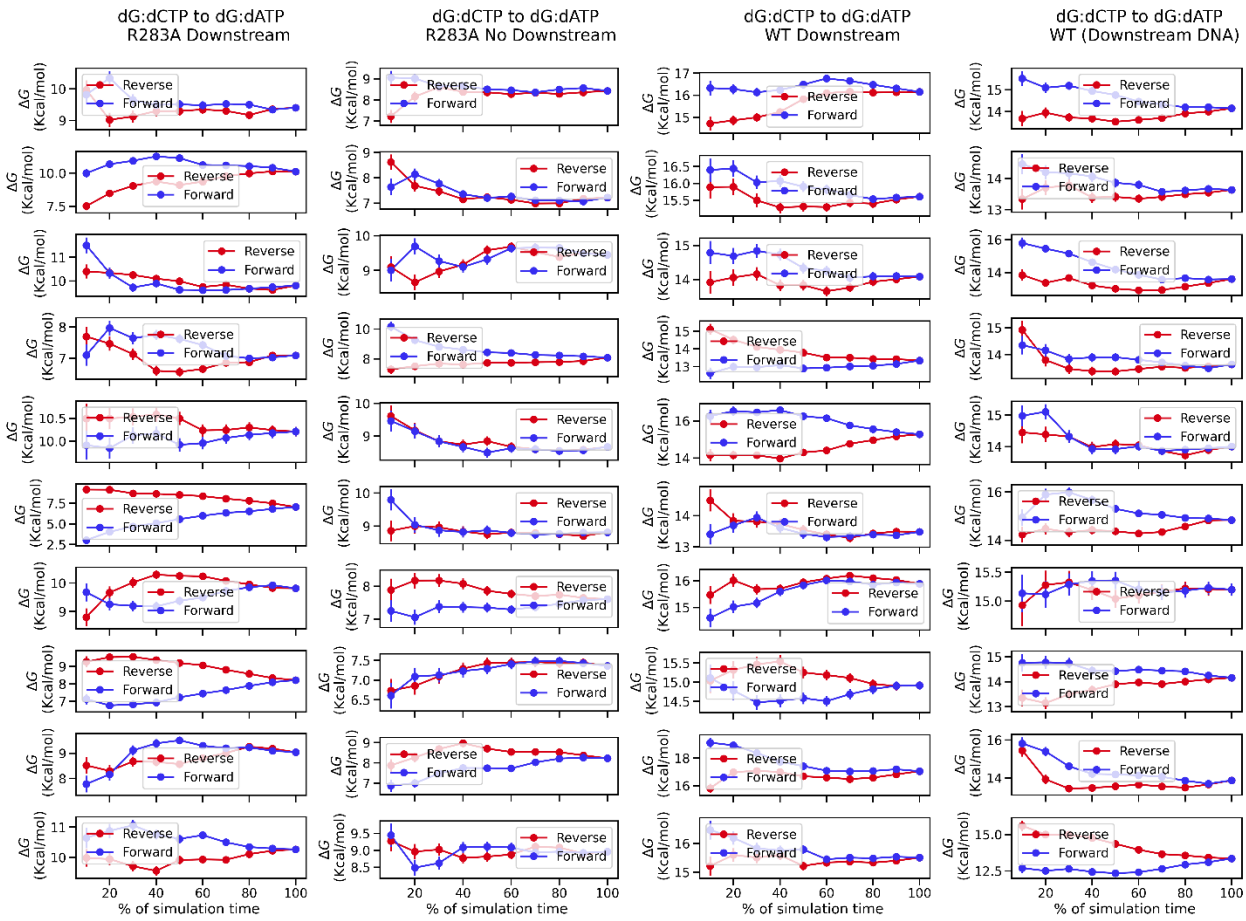


Figure S 15. Time evolution of the free energies (with error bars) for all the studied systems.

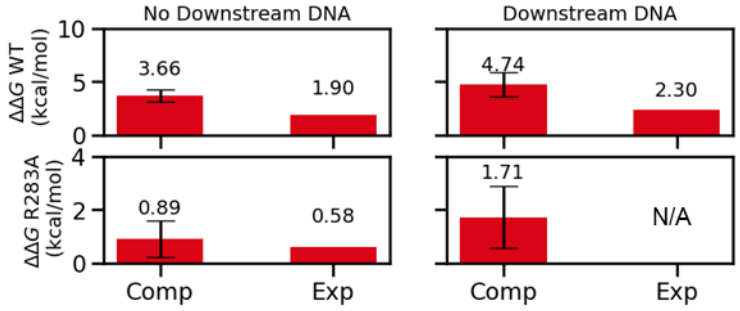


Figure S16. Estimates of relative binding free energy for WT and R283A systems in presence/absence of Downstream DNA strand

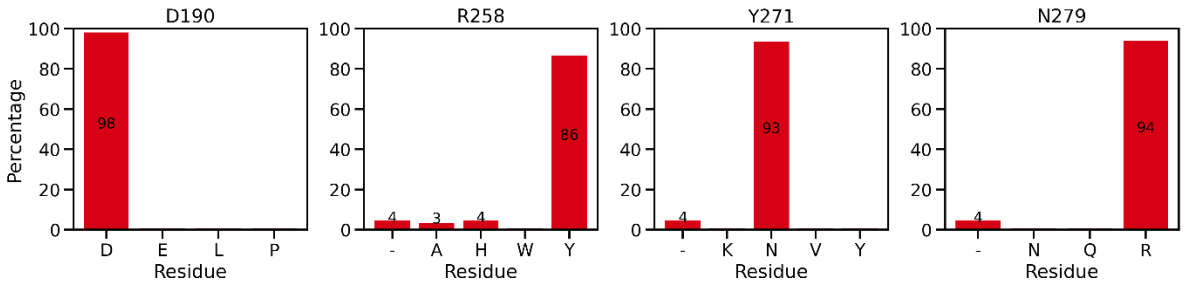


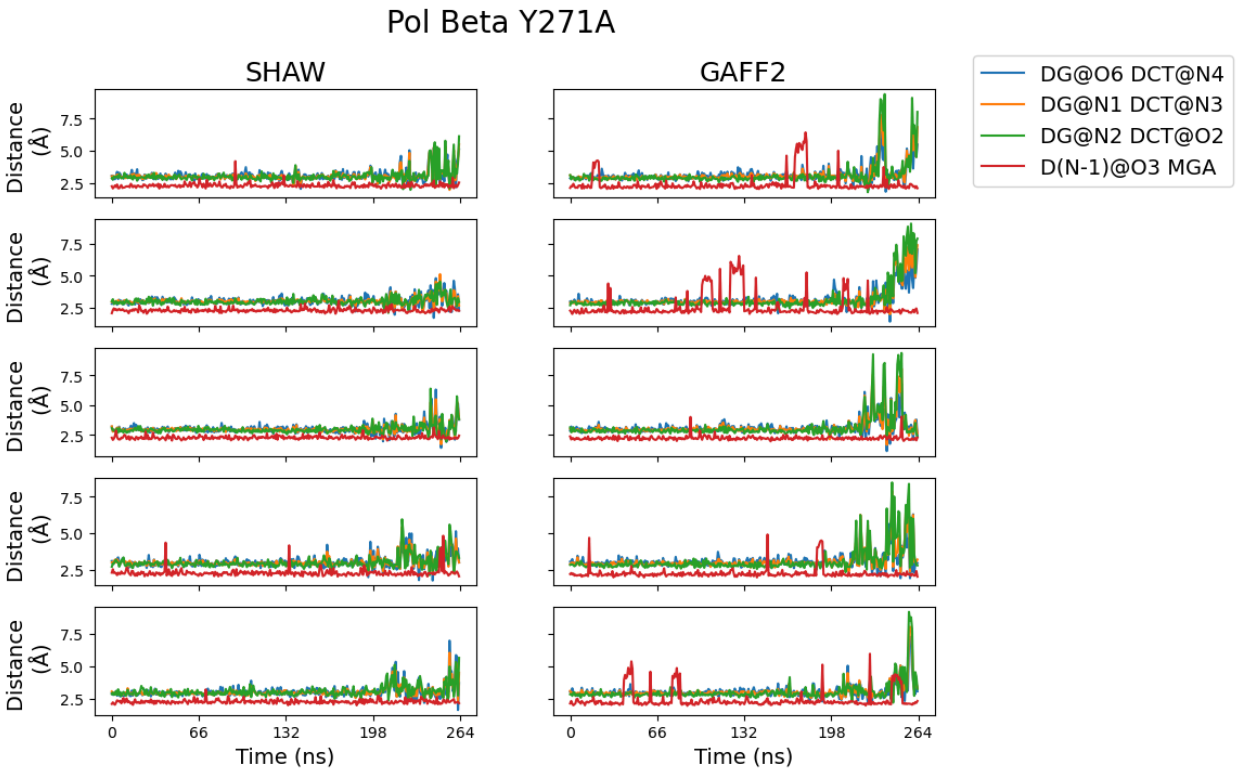
Figure S 17 Conservation percentages of discussed residues of MSA performed with consurf.

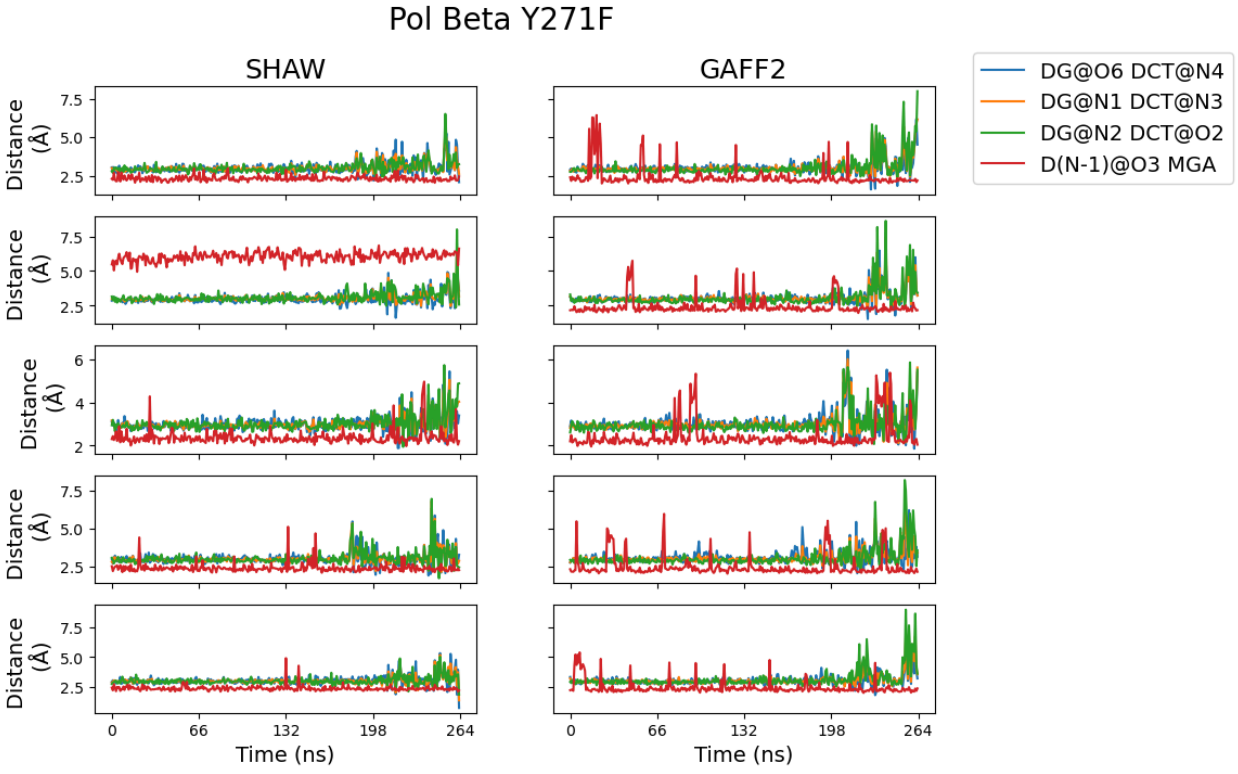
7.2. Chapter 4 Supplementary information:

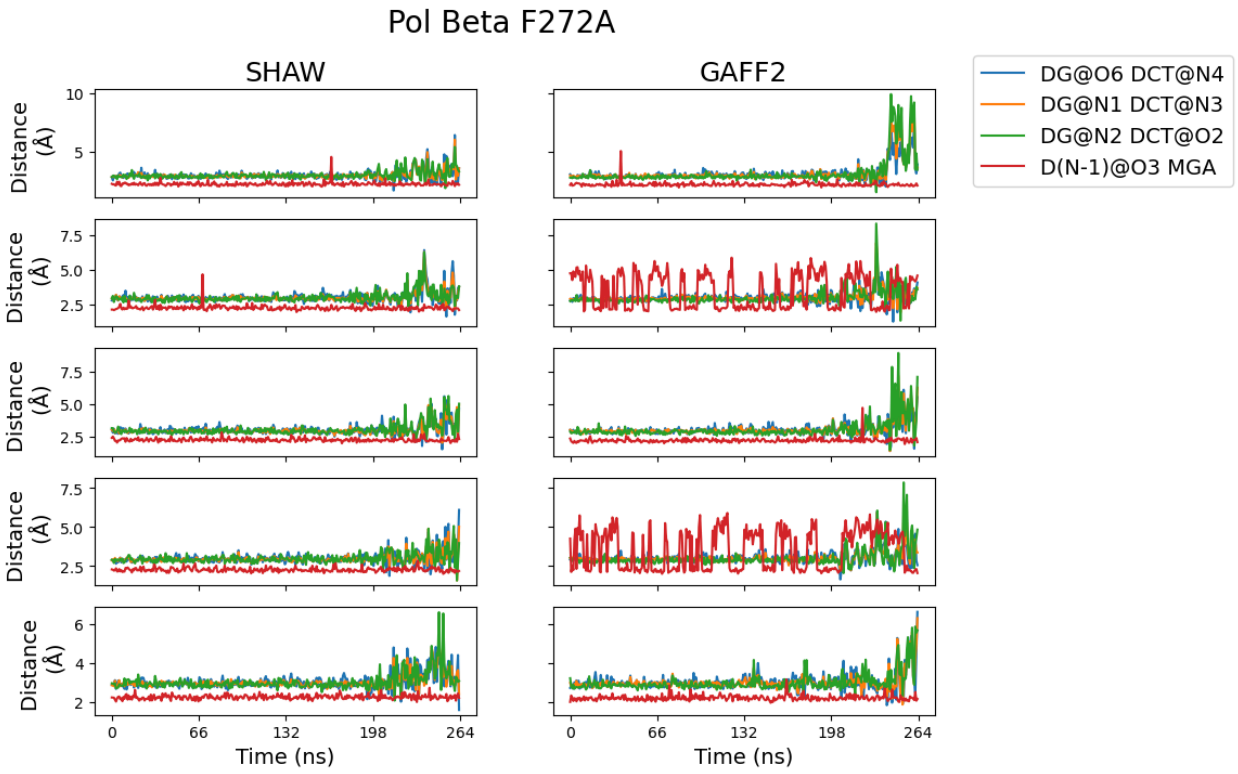
Table S 4. lambda scheduling and softcore parameters used in all transformations

λ	λ1	λ2	α	u0	w0
0	0	0	0.1	110	0
0.05	0	0.1	0.1	110	0
0.1	0	0.2	0.1	110	0
0.15	0	0.3	0.1	110	0
0.2	0	0.4	0.1	110	0
0.25	0	0.5	0.1	110	0
0.3	0.1	0.5	0.1	110	0

0.35	0.2	0.5	0.1	110	0
0.4	0.3	0.5	0.1	110	0
0.45	0.4	0.5	0.1	110	0
0.5	0.5	0.5	0.1	110	0


Pol Beta Y271A
SHAW
GAFF2
DG@O6 DCT@N4
DG@N1 DCT@N3
DG@N2 DCT@O2
D(N-1)@O3 MGA
Distance (Å)
Time (ns)


Pol Beta Y271F
SHAW
GAFF2
DG@O6 DCT@N4
DG@N1 DCT@N3
DG@N2 DCT@O2
D(N-1)@O3 MGA
Distance (Å)
Time (ns)


Pol Beta F272A
SHAW
GAFF2
DG@O6 DCT@N4
DG@N1 DCT@N3
DG@N2 DCT@O2
D(N-1)@O3 MGA
Distance (Å)
Time (ns)
0
66
132
198
264

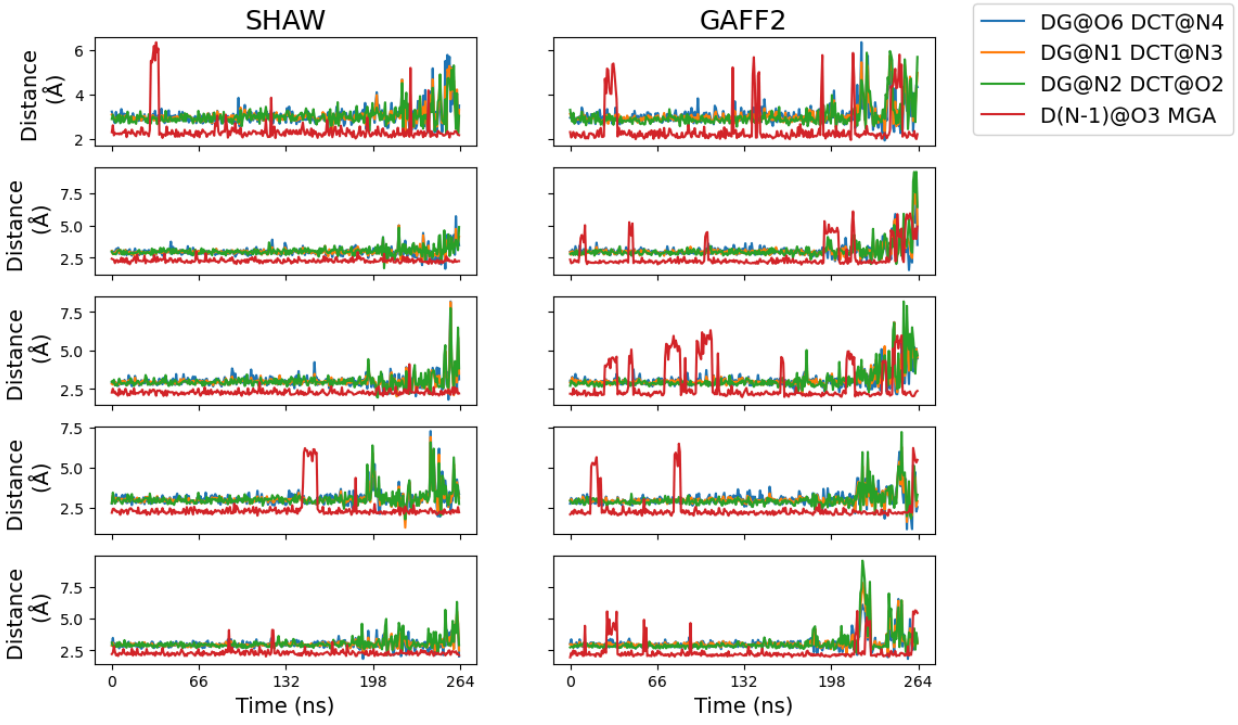
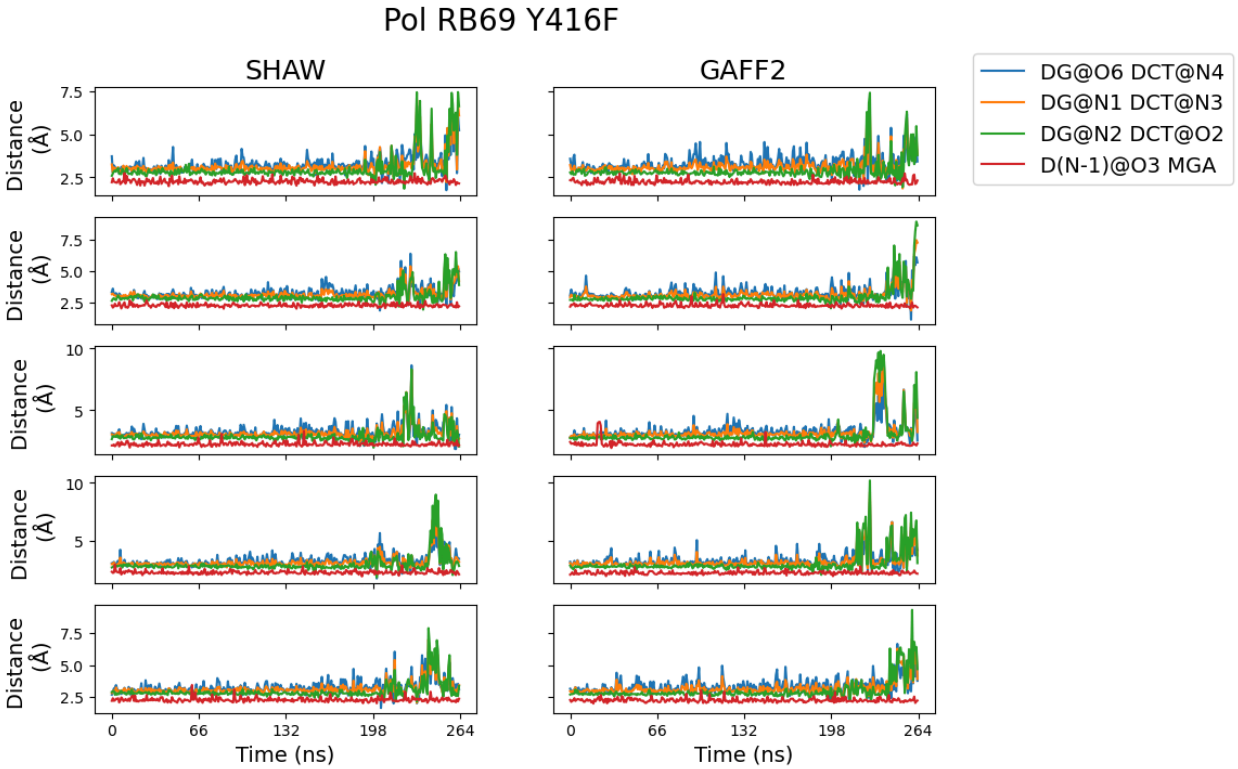
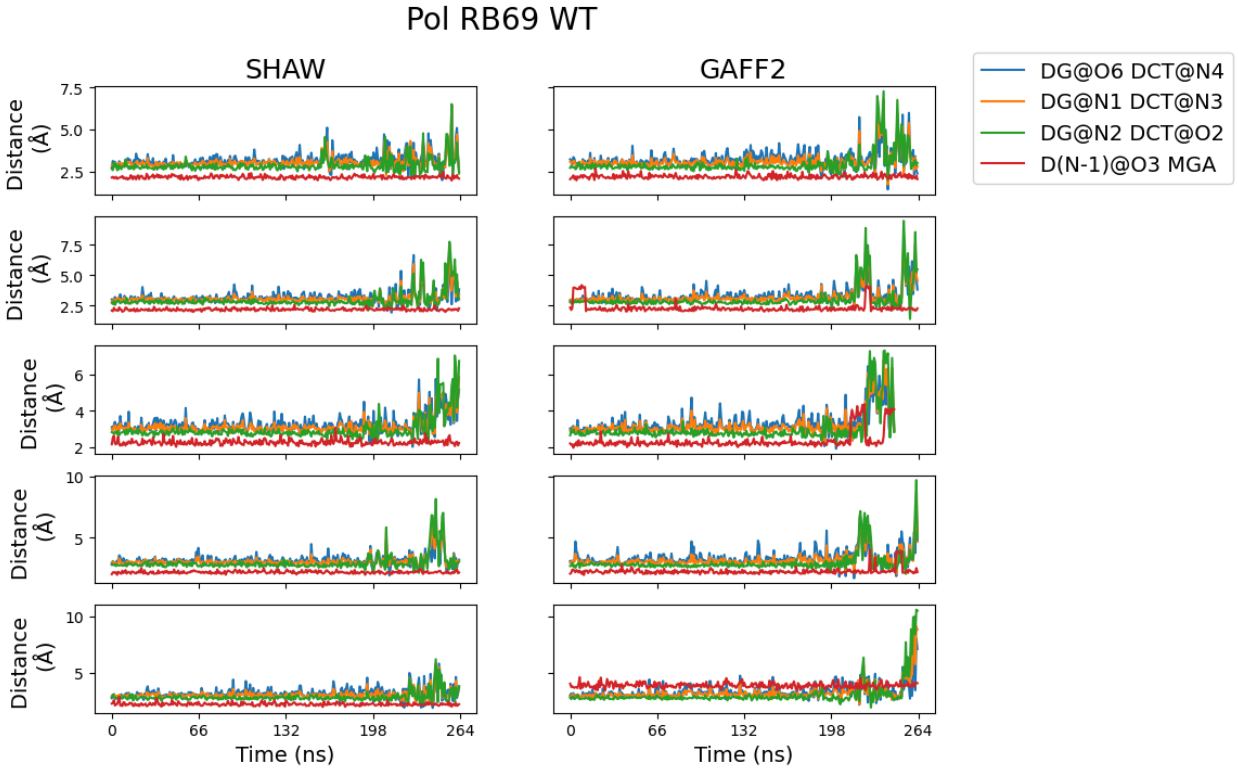

Pol Beta WT
SHAW
GAFF2
DG@O6 DCT@N4
DG@N1 DCT@N3
DG@N2 DCT@O2
D(N-1)@O3 MGA
Distance (Å)
Time (ns)
0
66
132
198
264

Figure S 18. Comparison of relevant distances in both SHAW and GAFF2 transformations for dCTP to rCTP in Polymerase Beta WT and mutants.


Pol RB69 Y416F
SHAW
GAFF2
DG@O6 DCT@N4
DG@N1 DCT@N3
DG@N2 DCT@O2
D(N-1)@O3 MGA
Distance (Å)
Time (ns)


Pol RB69 WT
SHAW
GAFF2
DG@O6 DCT@N4
DG@N1 DCT@N3
DG@N2 DCT@O2
D(N-1)@O3 MGA
Distance (Å)
Time (ns)

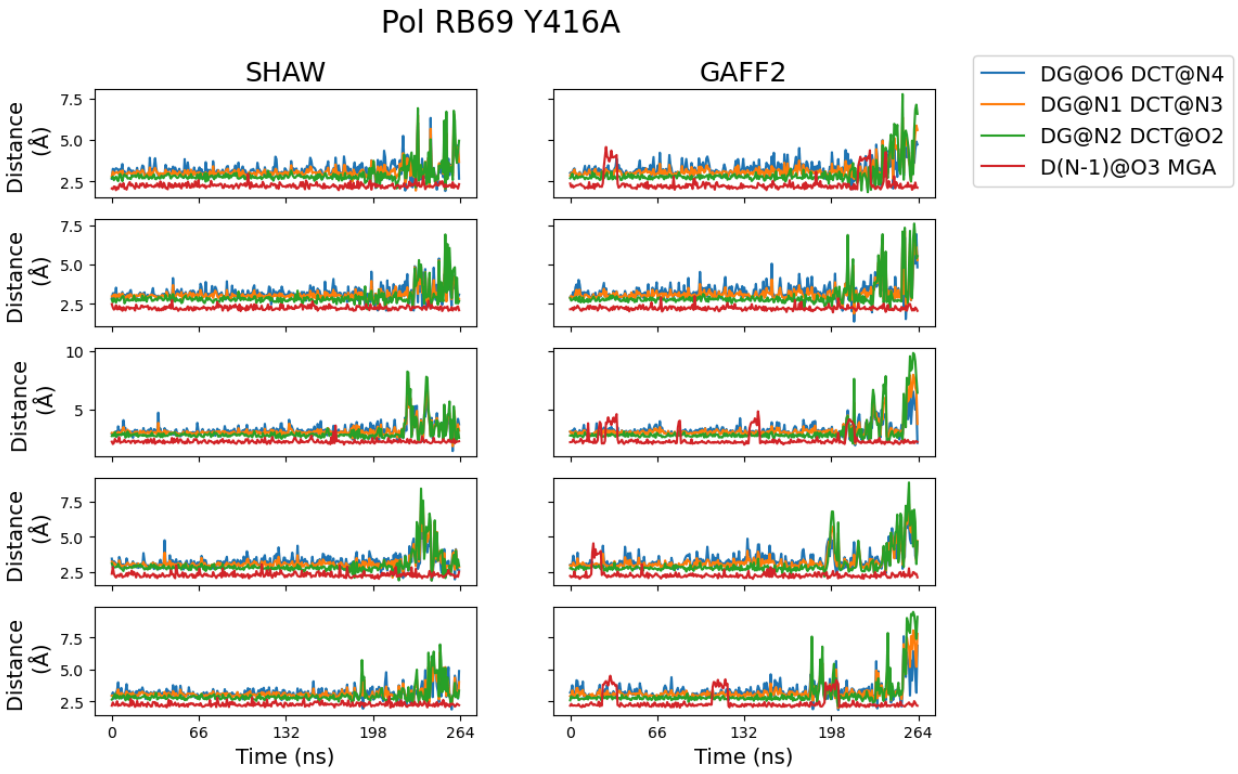


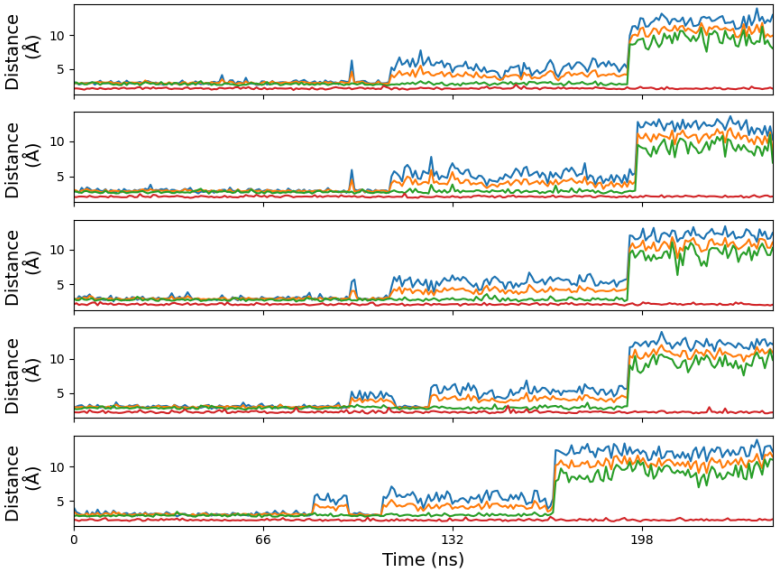
Figure S 19. Comparison of relevant distances in both SHAW and GAFF2 transformations for dCTP to rCTP in Polymerase RB69.

Table S 5 Average distances and s.d for all the transformation from dCTP to rCTP in both Pol Beta and Pol RB69

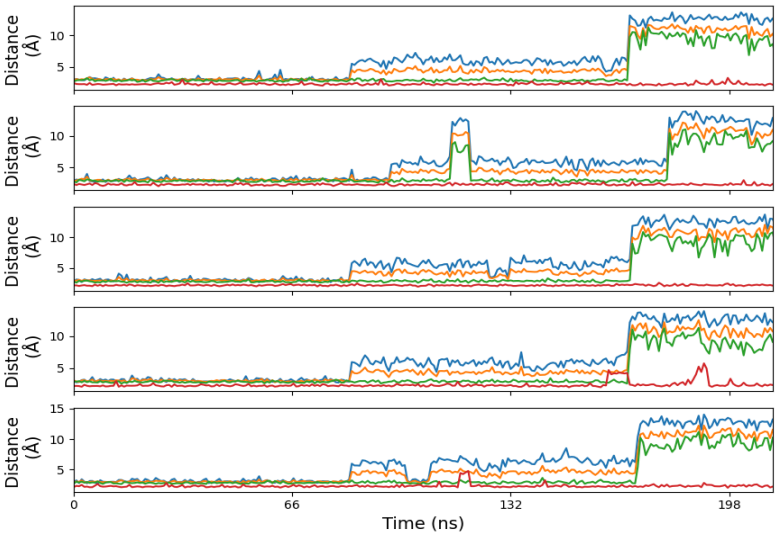
Pol	FF	Mutant	Distance	Mean	Std
Pol Beta	GAFF2	F272A	D(N-1)@O3 MGA	2.72	1.02
			DG@N1 DCT@N3	3.07	0.59
			DG@N2 DCT@O2	3.06	0.76
			DG@O6 DCT@N4	3.07	0.52
		WT	D(N-1)@O3 MGA	2.53	0.91
			DG@N1 DCT@N3	3.08	0.57
			DG@N2 DCT@O2	3.07	0.71
			DG@O6 DCT@N4	3.07	0.52
		Y271A	D(N-1)@O3 MGA	2.36	0.62
			DG@N1 DCT@N3	3.13	0.72
			DG@N2 DCT@O2	3.11	0.95
			DG@O6 DCT@N4	3.11	0.59
		Y271F	D(N-1)@O3 MGA	2.40	0.61

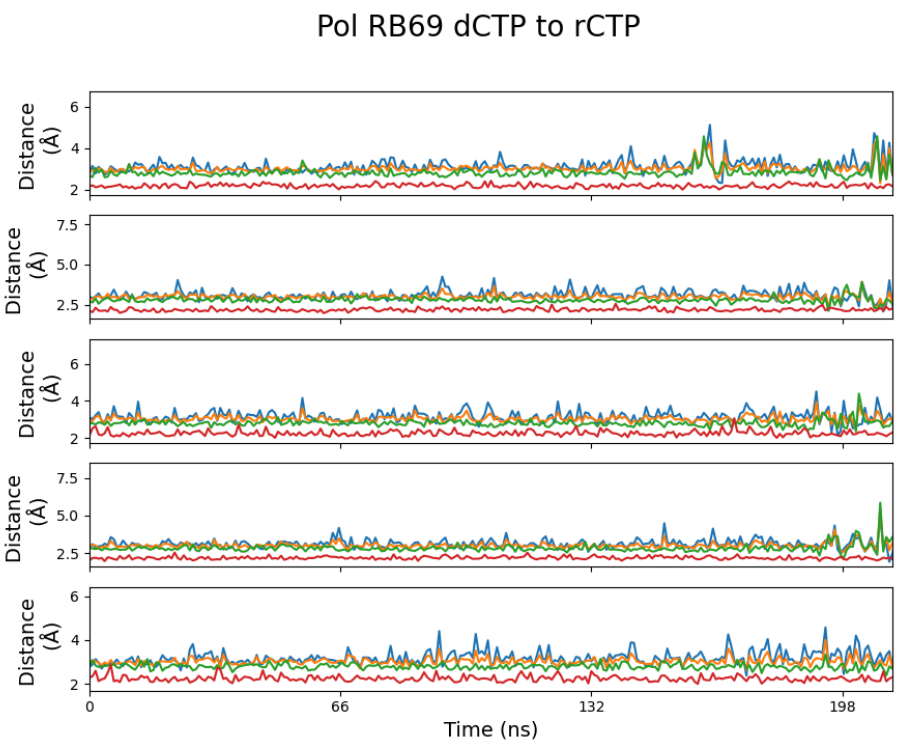
			DG@N1 DCT@N3	3.05	0.48
			DG@N2 DCT@O2	3.05	0.65
			DG@O6 DCT@N4	3.05	0.46
	SHAW	**F272A**	**D(N-1)@O3 MGA**	2.26	0.15
			DG@N1 DCT@N3	3.01	0.36
			DG@N2 DCT@O2	3.02	0.41
			DG@O6 DCT@N4	3.03	0.43
		WT	**D(N-1)@O3 MGA**	2.34	0.45
			DG@N1 DCT@N3	3.02	0.39
			DG@N2 DCT@O2	3.04	0.43
			DG@O6 DCT@N4	3.02	0.44
		Y271A	**D(N-1)@O3 MGA**	2.28	0.18
			DG@N1 DCT@N3	3.03	0.37
			DG@N2 DCT@O2	3.01	0.40
			DG@O6 DCT@N4	3.04	0.43
		Y271F	**D(N-1)@O3 MGA**	3.09	1.50
			DG@N1 DCT@N3	3.01	0.34
			DG@N2 DCT@O2	3.03	0.40
			DG@O6 DCT@N4	3.00	0.41
Pol RB69	**GAFF 2**	**WT**	**D(N-1)@O3 MGA**	2.61	0.73
			DG@N1 DCT@N3	3.23	0.68
			DG@N2 DCT@O2	3.02	0.89
			DG@O6 DCT@N4	3.32	0.63
		Y416A	**D(N-1)@O3 MGA**	2.34	0.45
			DG@N1 DCT@N3	3.24	0.72
			DG@N2 DCT@O2	3.05	1.00
			DG@O6 DCT@N4	3.31	0.62
		Y416F	**D(N-1)@O3 MGA**	2.25	0.16
			DG@N1 DCT@N3	3.21	0.66
			DG@N2 DCT@O2	3.00	0.88
			DG@O6 DCT@N4	3.31	0.61
	SHAW	**WT**	**D(N-1)@O3 MGA**	2.21	0.12
			DG@N1 DCT@N3	3.11	0.47
			DG@N2 DCT@O2	2.93	0.59
			DG@O6 DCT@N4	3.20	0.50
		Y416A	**D(N-1)@O3 MGA**	2.26	0.14
			DG@N1 DCT@N3	3.13	0.52
			DG@N2 DCT@O2	2.95	0.65
			DG@O6 DCT@N4	3.21	0.52
		Y416F	**D(N-1)@O3 MGA**	2.27	0.15
			DG@N1 DCT@N3	3.16	0.55
			DG@N2 DCT@O2	2.96	0.68
			DG@O6 DCT@N4	3.27	0.56

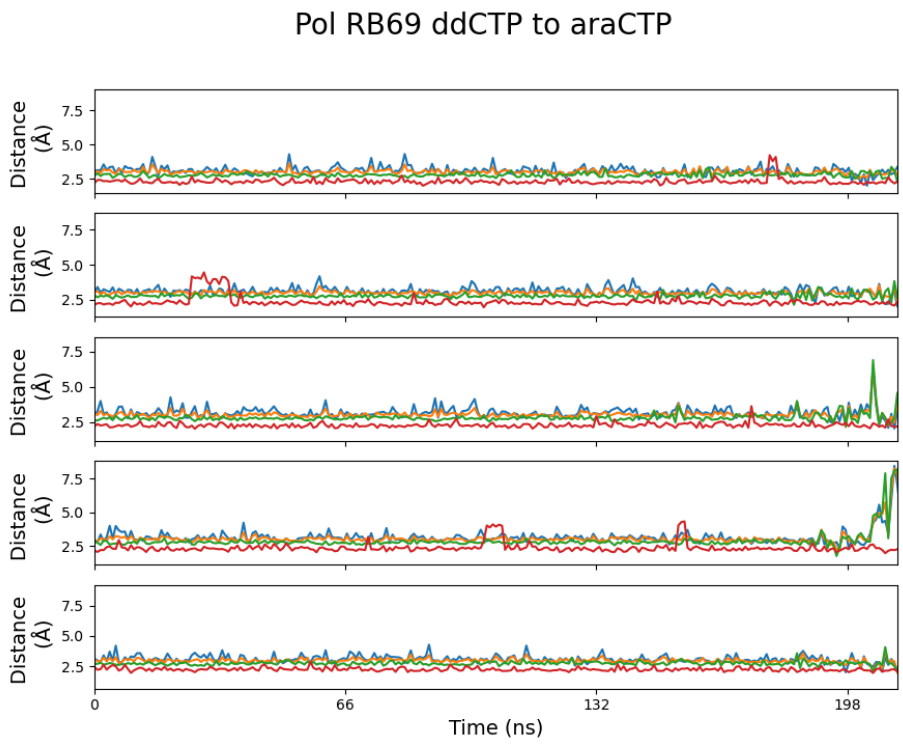
Pol RB69 dCTP to araCTP



Pol RB69 dCTP to ddCTP




Pol RB69 dCTP to rCTP
Distance (Å)
Distance (Å)
Distance (Å)
Distance (Å)
Distance (Å)
Time (ns)
0
66
132
198



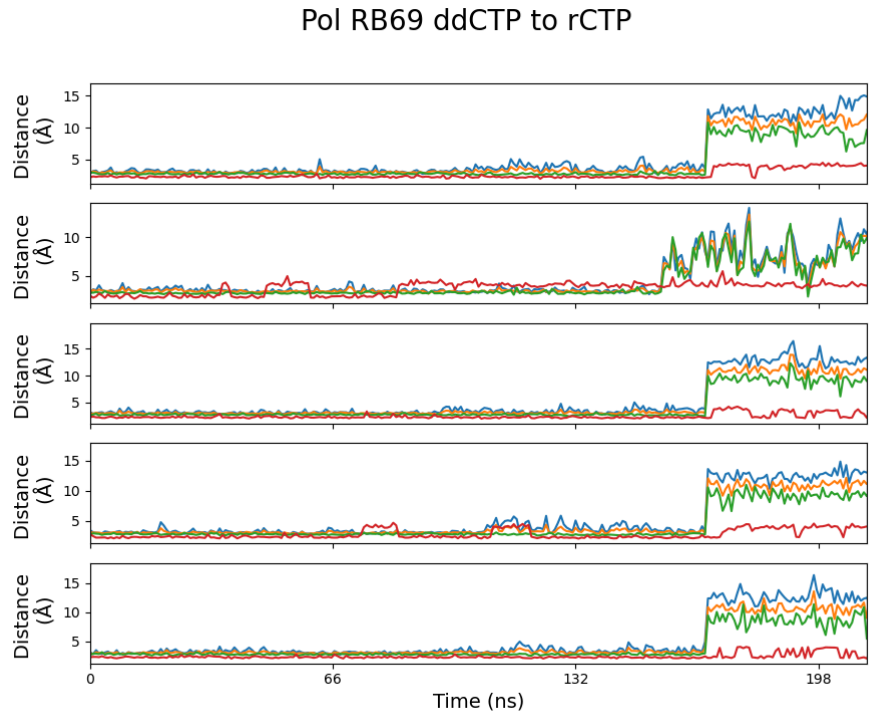


Figure S 20. Time series for relevant geometrical parameters in transformations involving dCTP,ddCTP,rCTP and araCTP in Pol RB69.

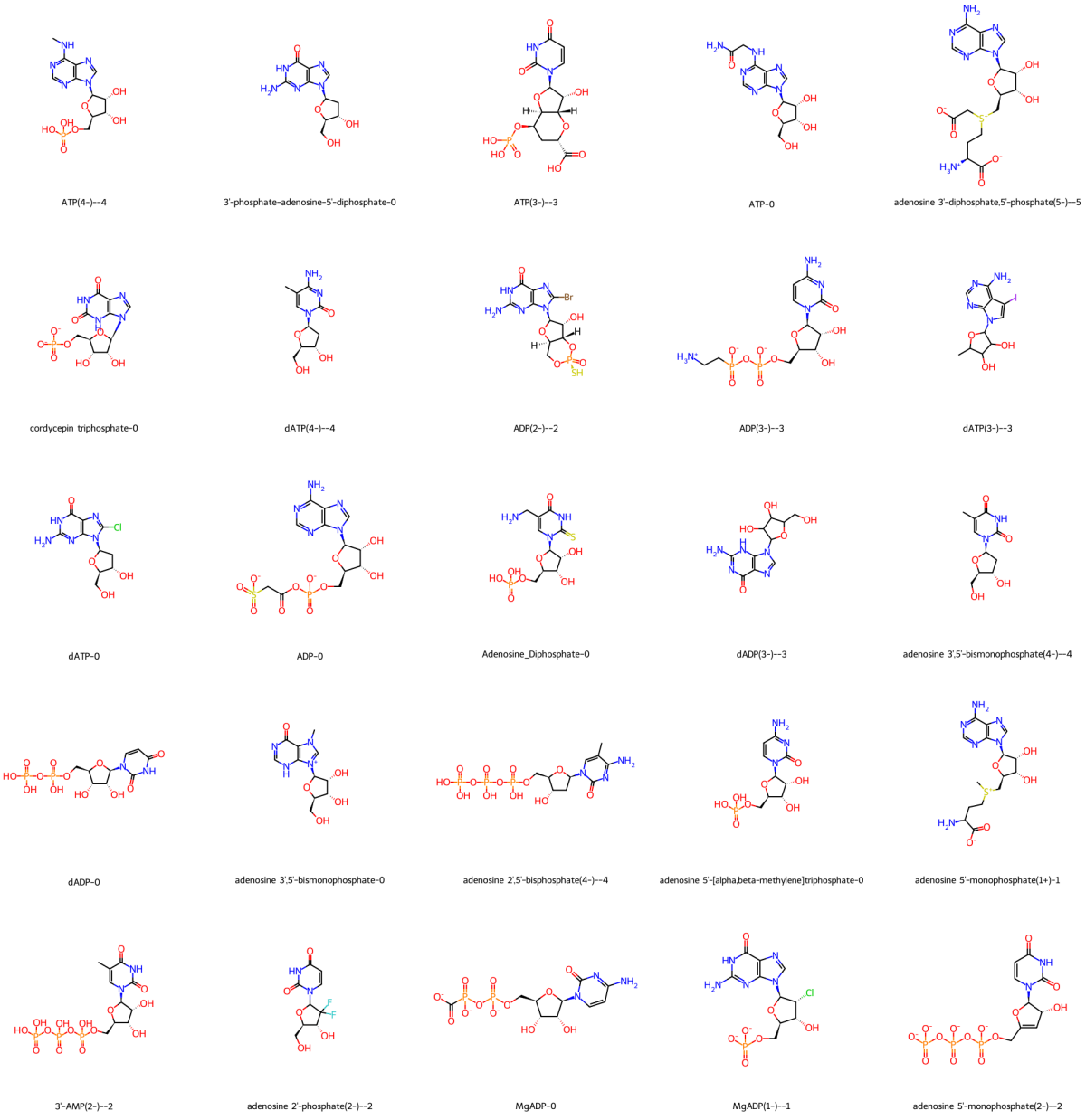


Figure S 21. Example of nucleotide modifications included in the NNP dataset.

7.3. Chapter 5 Supplementary information.

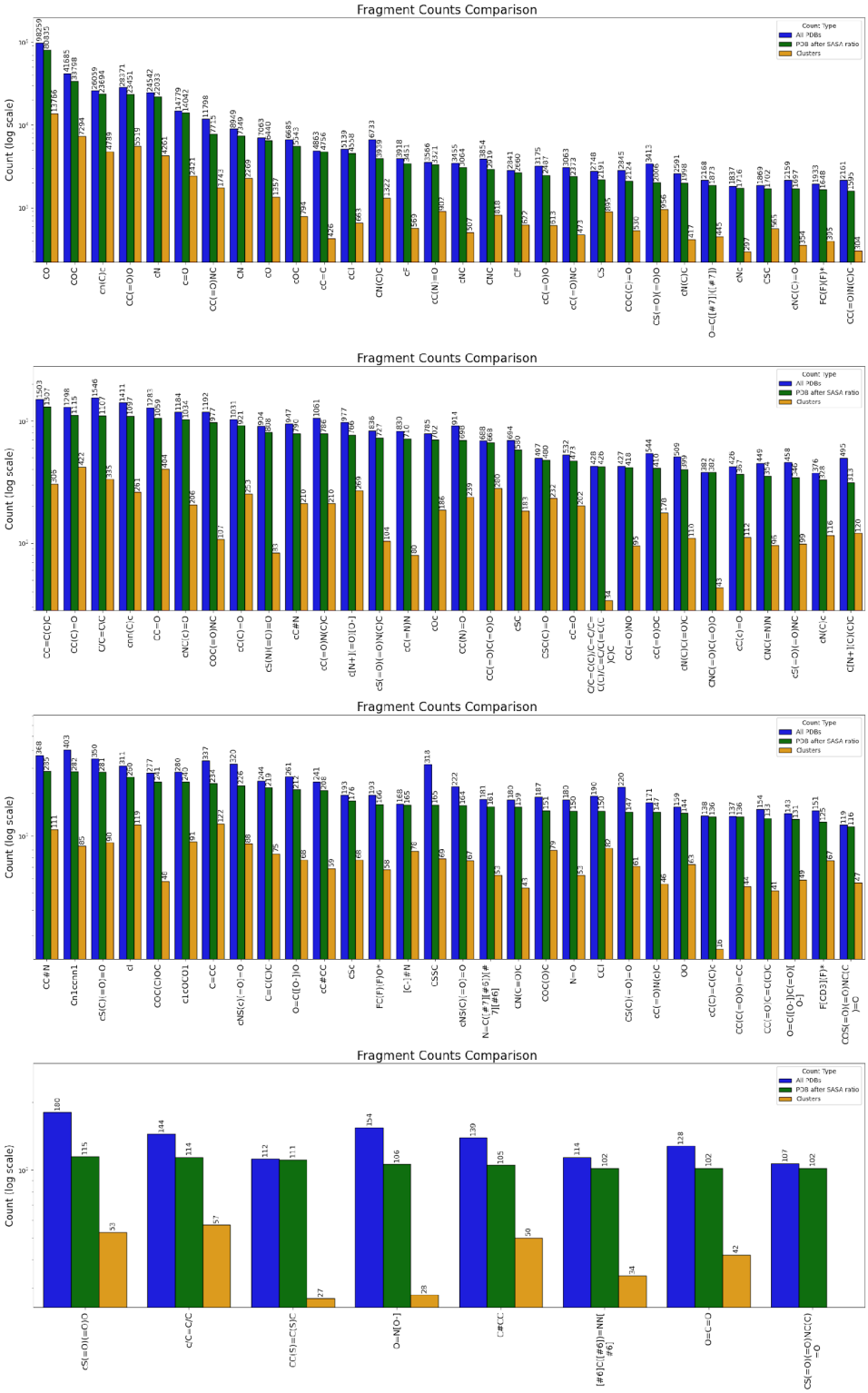


Figure S 22. Fragment distribution analysis across filtering stages for Ertl functional groups and Murcko scaffolds. Bar plots showing fragment counts at three stages: initial PDB entries (blue), after SASA ratio filtering < 0.25 for Ertl and < 0.5 for Murcko (green), and after protein sequence clustering at 50% identity (orange). Top three panels display Ertl fragmentation results

for heteroatom-containing functional groups, with notation distinguishing aromatic (lowercase 'c') from aliphatic (uppercase 'C') carbon attachments. Bottom panel shows Murcko scaffold decomposition results for ring systems and core structures. The progressive reduction from PDB to clusters demonstrates the elimination of redundancy while maintaining chemical diversity.

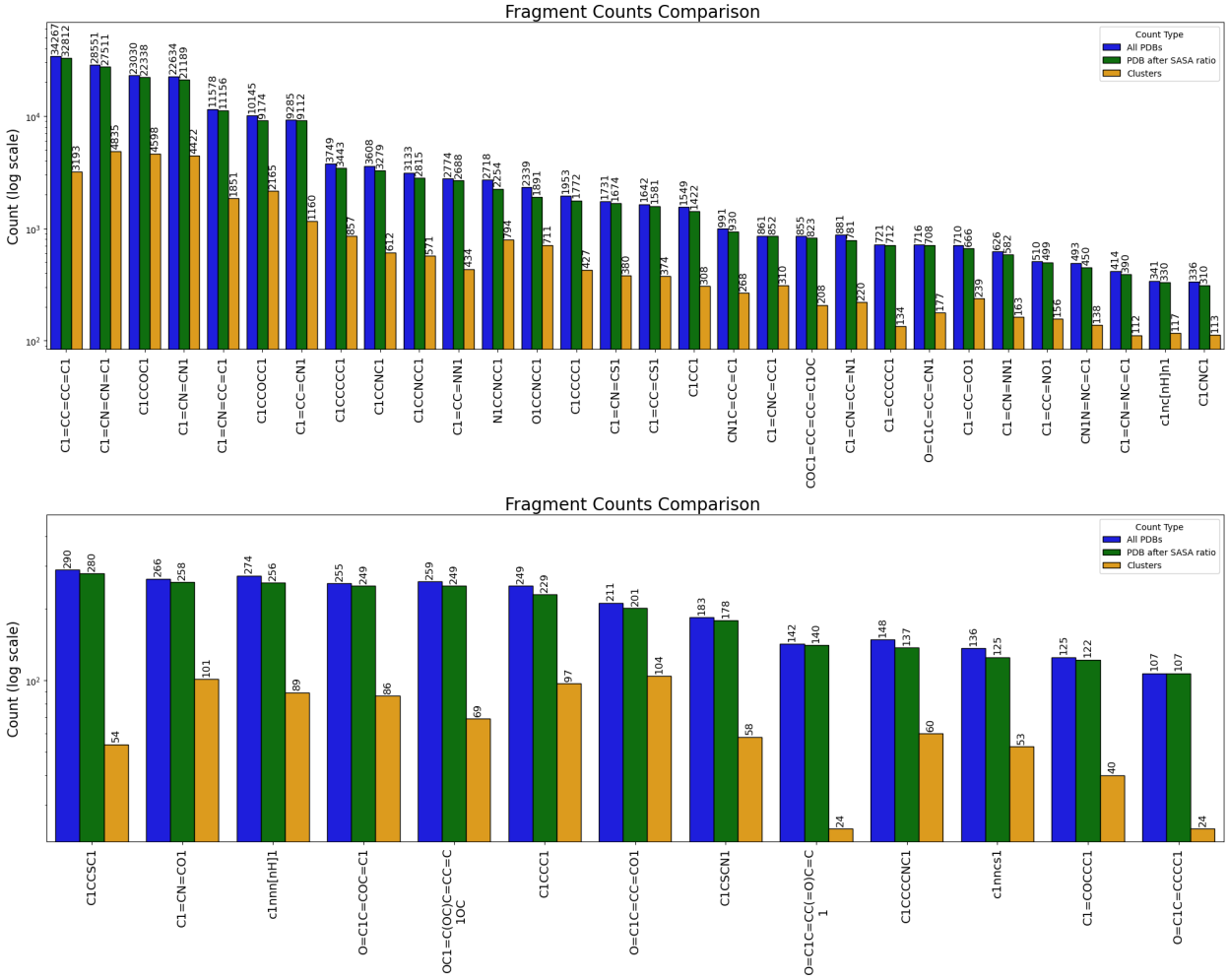


Figure S 23. Extended fragment distribution for additional Murcko scaffolds and complex heterocycles. Continuation of fragment count analysis showing less common but chemically important ring systems identified through Murcko decomposition, including fused rings, heterocyclic systems, and specialized scaffolds. The data follows the same three-stage filtering process: all PDBs (blue), post-SASA filtering (green), and post-clustering (orange), revealing that complex scaffolds generally show lower occurrence but maintain representation across diverse protein families after clustering.

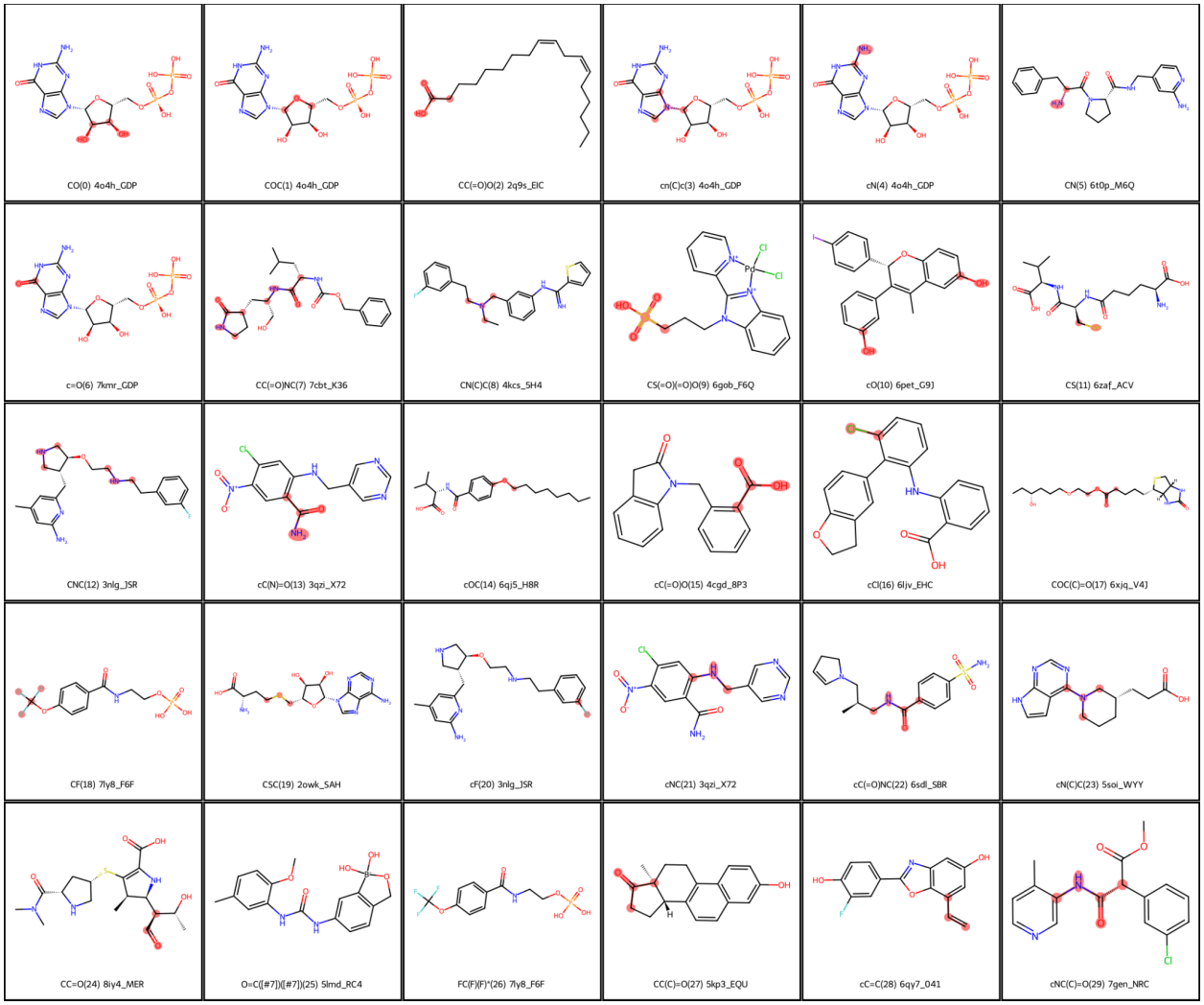


Figure S 24. 30 Representative examples of PDB ligands with highlighted Ertl functional group fragments. Each panel shows a co-crystallized ligand from the filtered PDB dataset with specific heteroatom-containing fragments highlighted in color, demonstrating the diversity of functional groups identified by the Ertl fragmentation algorithm.

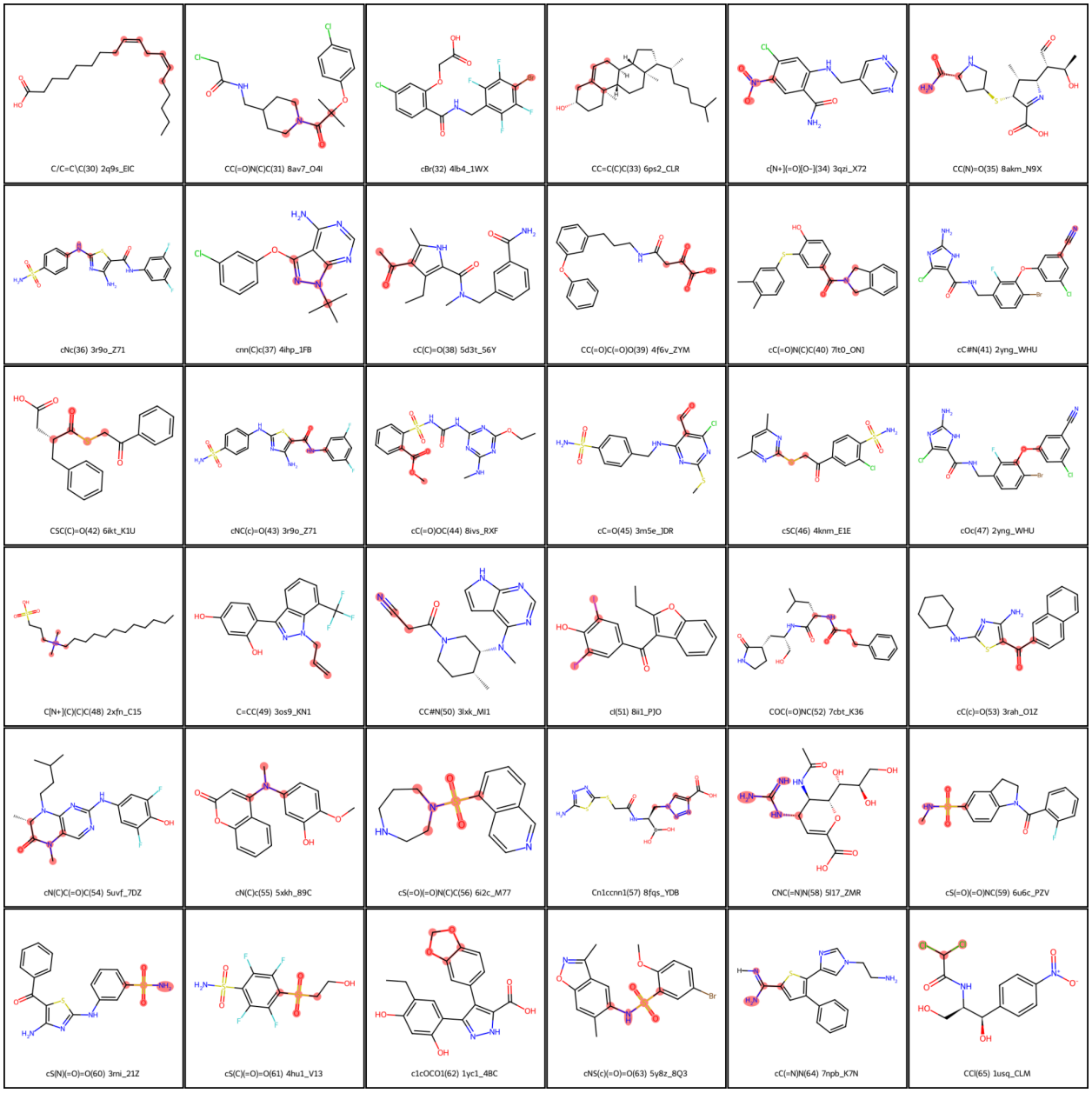


Figure S 25. 35 Representative examples of PDB ligands with highlighted Ertl functional group fragments. Each panel shows a co-crystallized ligand from the filtered PDB dataset with specific heteroatom-containing fragments highlighted in color, demonstrating the diversity of functional groups identified by the Ertl fragmentation algorithm.

8. REFERENCES

(1) Geronimo, I.; Vidossich, P.; Donati, E.; De Vivo, M. Computational Investigations of Polymerase Enzymes: Structure, Function, Inhibition, and Biotechnology. *WIREs Comput. Mol. Sci.* **2021**, *11* (6), e1534. https://doi.org/10.1002/wcms.1534.

(2) Wu, W.-J.; Yang, W.; Tsai, M.-D. How DNA Polymerases Catalyse Replication and Repair with Contrasting Fidelity. **2017**, *1* (9), 1–16. https://doi.org/10.1038/s41570-017-0068.

(3) Bebenek, K.; Kunkel, T. A. Functions of DNA Polymerases. In *Advances in Protein Chemistry*; Elsevier, 2004; Vol. 69, pp 137–165. https://doi.org/10.1016/S0065-3233(04)69005-X.

(4) Berdis, A. J. DNA Polymerases as Therapeutic Targets. *Biochemistry* **2008**, *47* (32), 8253–8260. https://doi.org/10.1021/bi801179f.

(5) Berdis, A. J. Inhibiting DNA Polymerases as a Therapeutic Intervention against Cancer. *Front. Mol. Biosci.* **2017**, *4*. https://doi.org/10.3389/fmolb.2017.00078.

(6) Matute, R. A.; Yoon, H.; Warshel, A. Exploring the Mechanism of DNA Polymerases by Analyzing the Effect of Mutations of Active Site Acidic Groups in Polymerase β. *Proteins Struct. Funct. Bioinforma.* **2016**, *84* (11), 1644–1657. https://doi.org/10.1002/prot.25106.

(7) Batra, V. K.; Beard, W. A.; Shock, D. D.; Krahn, J. M.; Pedersen, L. C.; Wilson, S. H. Magnesium-Induced Assembly of a Complete DNA Polymerase Catalytic Complex. *Structure* **2006**, *14* (4), 757–766. https://doi.org/10.1016/j.str.2006.01.011.

(8) Chang, C.; Zhou, G.; Gao, Y. In Crystallo Observation of Active Site Dynamics and Transient Metal Ion Binding within DNA Polymerases. *Struct. Dyn.* **2023**, *10* (3), 034702. https://doi.org/10.1063/4.0000187.

(9) Genna, V.; Vidossich, P.; Vidossich, P.; Ippoliti, E.; Carloni, P.; De Vivo, M. A Self-Activated Mechanism for Nucleic Acid Polymerization Catalyzed by DNA/RNA Polymerases. *J. Am. Chem. Soc.* **2016**, *138* (44), 14592–14598. https://doi.org/10.1021/jacs.6b05475.

(10) Geronimo, I.; Vidossich, P.; De Vivo, M. Local Structural Dynamics at the Metal-Centered Catalytic Site of Polymerases Is Critical for Fidelity. *ACS Catal.* **2021**, 14110–14121. https://doi.org/10.1021/acscatal.1c03840.

(11) Shankar, S.; Pan, J.; Yang, P.; Bian, Y.; Oroszlán, G.; Yu, Z.; Mukherjee, P.; Filman, D. J.; Hogle, J. M.; Shekhar, M.; Coen, D. M.; Abraham, J. Viral DNA Polymerase Structures Reveal Mechanisms of Antiviral Drug Resistance. *Cell* **2024**, *187* (20), 5572-5586.e15. https://doi.org/10.1016/j.cell.2024.07.048.

(12) Gardner, A. F.; Kelman, Z. DNA Polymerases in Biotechnology. *Front. Microbiol.* **2014**, *5*, 659. https://doi.org/10.3389/fmicb.2014.00659.

(13) Gahlon, H. L.; Sturla, S. J. Determining Steady-State Kinetics of DNA Polymerase Nucleotide Incorporation. In *Non-Natural Nucleic Acids: Methods and Protocols*; Shank, N., Ed.; Methods in Molecular Biology; Springer: New York, NY, 2019; pp 299–311. https://doi.org/10.1007/978-1-4939-9216-4_19.

(14) Czernecki, D.; Nourisson, A.; Legrand, P.; Delarue, M. Reclassification of Family A DNA Polymerases Reveals Novel Functional Subfamilies and Distinctive Structural Features. *Nucleic Acids Res.* **2023**, *51* (9), 4488–4507. https://doi.org/10.1093/nar/gkad242.
(15) Kazlauskas, D.; Krupovic, M.; Guglielmini, J.; Forterre, P.; Venclovas, Č. Diversity and Evolution of B-Family DNA Polymerases. *Nucleic Acids Res.* **2020**, *48* (18), 10142–10156. https://doi.org/10.1093/nar/gkaa760.
(16) Doublié, S.; Zahn, K. E. Structural Insights into Eukaryotic DNA Replication. *Front. Microbiol.* **2014**, *5*. https://doi.org/10.3389/fmicb.2014.00444.
(17) Beard, W. A.; Wilson, S. H. Structure and Mechanism of DNA Polymerase β. *Biochemistry* **2014**, *53* (17), 2768–2780. https://doi.org/10.1021/bi500139h.
(18) Yang, W. An Overview of Y-Family DNA Polymerases and a Case Study of Human DNA Polymerase η. *Biochemistry* **2014**, *53* (17), 2793–2803. https://doi.org/10.1021/bi500019s.
(19) Ghosh, D.; Raghavan, S. C. 20 Years of DNA Polymerase μ, the Polymerase That Still Surprises. *FEBS J.* **2021**, *288* (24), 7230–7242. https://doi.org/10.1111/febs.15852.
(20) Prindle, M. J.; Schmitt, M. W.; Parmeggiani, F.; Loeb, L. A. A Substitution in the Fingers Domain of DNA Polymerase δ Reduces Fidelity by Altering Nucleotide Discrimination in the Catalytic Site*. *J. Biol. Chem.* **2013**, *288* (8), 5572–5580. https://doi.org/10.1074/jbc.M112.436410.
(21) Arora, K.; Schlick, T. In Silico Evidence for DNA Polymerase-β's Substrate-Induced Conformational Change. *Biophys. J.* **2004**, *87* (5), 3088–3099. https://doi.org/10.1529/biophysj.104.040915.
(22) Schlick, T.; Arora, K.; Beard, W. A.; Wilson, S. H. Perspective: Pre-Chemistry Conformational Changes in DNA Polymerase Mechanisms. *Theor. Chem. Acc.* **2012**, *131* (12), 1287. https://doi.org/10.1007/s00214-012-1287-7.
(23) Yang, G.; Franklin, M.; Li, J.; Lin, T.-C.; Konigsberg, W. A Conserved Tyr Residue Is Required for Sugar Selectivity in a Pol α DNA Polymerase. *Biochemistry* **2002**, *41* (32), 10256–10261. https://doi.org/10.1021/bi0202171.
(24) Betancurt-Anzola, L.; Martínez-Carranza, M.; Delarue, M.; Zatopek, K. M.; Gardner, A. F.; Sauguet, L. Molecular Basis for Proofreading by the Unique Exonuclease Domain of Family-D DNA Polymerases. *Nat. Commun.* **2023**, *14* (1), 8306. https://doi.org/10.1038/s41467-023-44125-x.
(25) Brautigam, C. A.; Steitz, T. A. Structural Principles for the Inhibition of the 3′-5′ Exonuclease Activity of *Escherichia Coli* DNA Polymerase I by Phosphorothioates 1. *J. Mol. Biol.* **1998**, *277* (2), 363–377. https://doi.org/10.1006/jmbi.1997.1586.
(26) Wu, S.; Beard, W. A.; Pedersen, L. G.; Wilson, S. H. Structural Comparison of DNA Polymerase Architecture Suggests a Nucleotide Gateway to the Polymerase Active Site. *Chem. Rev.* **2014**, *114* (5), 2759–2774. https://doi.org/10.1021/cr3005179.
(27) Genna, V.; Carloni, P.; De Vivo, M. A Strategically Located Arg/Lys Residue Promotes Correct Base Paring During Nucleic Acid Biosynthesis in Polymerases. *J. Am. Chem. Soc.* **2018**, *140* (9), 3312–3321. https://doi.org/10.1021/jacs.7b12446.
(28) Genna, V.; Donati, E.; De Vivo, M. The Catalytic Mechanism of DNA and RNA Polymerases. *ACS Catal.* **2018**, *8* (12), 11103–11118. https://doi.org/10.1021/acscatal.8b03363.
(29) Ahn, J.; Werneburg, B. G.; Tsai, M.-D. DNA Polymerase β: Structure−Fidelity Relationship from Pre-Steady-State Kinetic Analyses of All Possible Correct and Incorrect

Base Pairs for Wild Type and R283A Mutant. *Biochemistry* **1997**, *36* (5), 1100–1107. https://doi.org/10.1021/bi961653o.
(30) Wang, M.; Xia, S.; Blaha, G.; Steitz, T. A.; Konigsberg, W. H.; Wang, J. Insights into Base Selectivity from the 1.8 Å Resolution Structure of an RB69 DNA Polymerase Ternary Complex. *Biochemistry* **2011**, *50* (4), 581–590. https://doi.org/10.1021/bi101192f.
(31) Xia, S.; Konigsberg, W. H. RB69 DNA Polymerase Structure, Kinetics, and Fidelity. *Biochemistry* **2014**, *53* (17), 2752–2767. https://doi.org/10.1021/bi4014215.
(32) Lee, H. R.; Wang, M.; Konigsberg, W. The Reopening Rate of the Fingers Domain Is a Determinant of Base Selectivity for RB69 DNA Polymerase. *Biochemistry* **2009**, *48* (10), 2087–2098. https://doi.org/10.1021/bi8016284.
(33) Gong, S.; Kirmizialtin, S.; Chang, A.; Mayfield, J. E.; Zhang, Y. J.; Johnson, K. A. Kinetic and Thermodynamic Analysis Defines Roles for Two Metal Ions in DNA Polymerase Specificity and Catalysis. *J. Biol. Chem.* **2021**, *296*, 100184. https://doi.org/10.1074/jbc.RA120.016489.
(34) Arora, K.; Beard, W. A.; Wilson, S. H.; Schlick, T. Mismatch-Induced Conformational Distortions in Polymerase β Support an Induced-Fit Mechanism for Fidelity. *Biochemistry* **2005**, *44* (40), 13328–13341. https://doi.org/10.1021/bi0507682.
(35) Klvaňa, M.; Murphy, D. L.; Jeřábek, P.; Goodman, M. F.; Warshel, A.; Sweasy, J. B.; Florián, J. Catalytic Effects of Mutations of Distant Protein Residues in Human DNA Polymerase β: Theory and Experiment. *Biochemistry* **2012**, *51* (44), 8829–8843. https://doi.org/10.1021/bi300783t.
(36) Dodd, T.; Botto, M.; Paul, F.; Fernandez-Leiro, R.; Lamers, M. H.; Ivanov, I. Polymerization and Editing Modes of a High-Fidelity DNA Polymerase Are Linked by a Well-Defined Path. *Nat. Commun.* **2020**, *11* (1), 5379. https://doi.org/10.1038/s41467-020-19165-2.
(37) Rucker, R.; Oelschlaeger, P.; Warshel, A. A Binding Free Energy Decomposition Approach for Accurate Calculations of the Fidelity of DNA Polymerases. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (3), 671–680. https://doi.org/10.1002/prot.22596.
(38) Cavanaugh, N. A.; Beard, W. A.; Wilson, S. H. DNA Polymerase β Ribonucleotide Discrimination. *J. Biol. Chem.* **2010**, *285* (32), 24457–24465. https://doi.org/10.1074/jbc.M110.132407.
(39) Smith, M. R.; Alnajjar, K. S.; Hoitsma, N. M.; Sweasy, J. B.; Freudenthal, B. D. Molecular and Structural Characterization of Oxidized Ribonucleotide Insertion into DNA by Human DNA Polymerase β. *J. Biol. Chem.* **2020**, *295* (6), 1613–1622. https://doi.org/10.1074/jbc.RA119.011569.
(40) Yoon, H.; Warshel, A. The Control of the Discrimination between dNTP and rNTP in DNA and RNA Polymerase. *Proteins Struct. Funct. Bioinforma.* **2016**, *84* (11), 1616–1624. https://doi.org/10.1002/prot.25104.
(41) Florián, J.; Goodman, M. F.; Warshel, A. Computer Simulation Studies of the Fidelity of DNA Polymerases. *Biopolymers* **2003**, *68* (3), 286–299. https://doi.org/10.1002/bip.10244.
(42) Galvelis, R.; Varela-Rial, A.; Doerr, S.; Fino, R.; Eastman, P.; Markland, T. E.; Chodera, J. D.; De Fabritiis, G. NNP/MM: Accelerating Molecular Dynamics Simulations with Machine Learning Potentials and Molecular Mechanics. *J. Chem. Inf. Model.* **2023**, *63* (18), 5701–5708. https://doi.org/10.1021/acs.jcim.3c00773.

(43) Sabanés Zariquiey, F.; Galvelis, R.; Gallicchio, E.; Chodera, J. D.; Markland, T. E.; De Fabritiis, G. Enhancing Protein–Ligand Binding Affinity Predictions Using Neural Network Potentials. *J. Chem. Inf. Model.* **2024**, *64* (5), 1481–1485. https://doi.org/10.1021/acs.jcim.3c02031.
(44) Cranford, M. T.; Kaszubowski, J. D.; Trakselis, M. A. A Hand-off of DNA between Archaeal Polymerases Allows High-Fidelity Replication to Resume at a Discrete Intermediate Three Bases Past 8-Oxoguanine. *Nucleic Acids Res.* **2020**, *48* (19), 10986–10997. https://doi.org/10.1093/nar/gkaa803.
(45) Anand, J.; Chiou, L.; Sciandra, C.; Zhang, X.; Hong, J.; Wu, D.; Zhou, P.; Vaziri, C. Roles of Trans-Lesion Synthesis (TLS) DNA Polymerases in Tumorigenesis and Cancer Therapy. *NAR Cancer* **2023**, *5* (1), zcad005. https://doi.org/10.1093/narcan/zcad005.
(46) Paniagua, I.; Jacobs, J. J. L. Freedom to Err: The Expanding Cellular Functions of Translesion DNA Polymerases. *Mol. Cell* **2023**, *83* (20), 3608–3621. https://doi.org/10.1016/j.molcel.2023.07.008.
(47) Dianov, G. L.; Hübscher, U. Mammalian Base Excision Repair: The Forgotten Archangel. *Nucleic Acids Res.* **2013**, *41* (6), 3483–3490. https://doi.org/10.1093/nar/gkt076.
(48) Fairlamb, M. S.; Spies, M.; Washington, M. T.; Freudenthal, B. D. Visualizing the Coordination of Apurinic/Apyrimidinic Endonuclease (APE1) and DNA Polymerase β during Base Excision Repair. *J. Biol. Chem.* **2023**, *299* (5), 104636. https://doi.org/10.1016/j.jbc.2023.104636.
(49) Howard, M. J.; Rodriguez, Y.; Wilson, S. H. DNA Polymerase β Uses Its Lyase Domain in a Processive Search for DNA Damage. *Nucleic Acids Res.* **2017**, *45* (7), 3822–3832. https://doi.org/10.1093/nar/gkx047.
(50) Hegde, M. L.; Hazra, T. K.; Mitra, S. Early Steps in the DNA Base Excision/Single-Strand Interruption Repair Pathway in Mammalian Cells. *Cell Res.* **2008**, *18* (1), 27–47. https://doi.org/10.1038/cr.2008.8.
(51) Kamzeeva, P. N.; Aralov, A. V.; Alferova, V. A.; Korshun, V. A. Recent Advances in Molecular Mechanisms of Nucleoside Antivirals. *Curr. Issues Mol. Biol.* **2023**, *45* (8), 6851–6879. https://doi.org/10.3390/cimb45080433.
(52) Langtry, H. D.; Campoli-Richards, D. M. Zidovudine. *Drugs* **1989**, *37* (4), 408–450. https://doi.org/10.2165/00003495-198937040-00003.
(53) Elion, G. B. The Biochemistry and Mechanism of Action of Acyclovir. *J. Antimicrob. Chemother.* **1983**, *12* (suppl_B), 9–17. https://doi.org/10.1093/jac/12.suppl_B.9.
(54) Iwasaki, H.; Huang, P.; Keating, M. J.; Plunkett, W. Differential Incorporation of Ara-C, Gemcitabine, and Fludarabine Into Replicating and Repairing DNA in Proliferating Human Leukemia Cells. *Blood* **1997**, *90* (1), 270–278. https://doi.org/10.1182/blood.V90.1.270.
(55) Öberg, B. Rational Design of Polymerase Inhibitors as Antiviral Drugs. *Antiviral Res.* **2006**, *71* (2), 90–95. https://doi.org/10.1016/j.antiviral.2006.05.012.
(56) Weberschock, T.; Gholam, P.; Hueter, E.; Flux, K.; Hartmann, M. Long-Term Efficacy and Safety of Once-Daily Nevirapine in Combination with Tenofovir and Emtricitabine in the Treatment of HIV-Infected Patients: A 72-Week Prospective Multicenter Study (TENOR-Trial). *Eur. J. Med. Res.* **2009**, *14* (12), 516–519. https://doi.org/10.1186/2047-783X-14-12-516.

(57) Kannan, S.; Gillespie, S. W.; Picking, W. L.; Picking, W. D.; Lorson, C. L.; Singh, K. Inhibitors against DNA Polymerase I Family of Enzymes: Novel Targets and Opportunities. *Biology* **2024**, *13* (4), 204. https://doi.org/10.3390/biology13040204.
(58) Qi, L.; Luo, Q.; Zhang, Y.; Jia, F.; Zhao, Y.; Wang, F. Advances in Toxicological Research of the Anticancer Drug Cisplatin. *Chem. Res. Toxicol.* **2019**, *32* (8), 1469–1486. https://doi.org/10.1021/acs.chemrestox.9b00204.
(59) Munafò, F.; Nigro, M.; Brindani, N.; Manigrasso, J.; Geronimo, I.; Ottonello, G.; Armirotti, A.; De Vivo, M. Computer-Aided Identification, Synthesis, and Biological Evaluation of DNA Polymerase H Inhibitors for the Treatment of Cancer. Social Science Research Network: Rochester, NY August 25, 2022. https://doi.org/10.2139/ssrn.4200126.
(60) Hottin, A.; Marx, A. Structural Insights into the Processing of Nucleobase-Modified Nucleotides by DNA Polymerases. *Acc. Chem. Res.* **2016**, *49* (3), 418–427. https://doi.org/10.1021/acs.accounts.5b00544.
(61) Chen, C.-Y. DNA Polymerases Drive DNA Sequencing-by-Synthesis Technologies: Both Past and Present. *Front. Microbiol.* **2014**, *5*. https://doi.org/10.3389/fmicb.2014.00305.
(62) Śpibida, M.; Krawczyk, B.; Olszewski, M.; Kur, J. Modified DNA Polymerases for PCR Troubleshooting. *J. Appl. Genet.* **2017**, *58* (1), 133–142. https://doi.org/10.1007/s13353-016-0371-4.
(63) Chowdhury, S.; Wang, J.; Nuccio, S. P.; Mao, H.; Di Antonio, M. Short LNA-Modified Oligonucleotide Probes as Efficient Disruptors of DNA G-Quadruplexes. *Nucleic Acids Res.* **2022**, *50* (13), 7247–7259. https://doi.org/10.1093/nar/gkac569.
(64) Malik, T. N.; Chaput, J. C. XNA Enzymes by Evolution and Design. *Curr. Res. Chem. Biol.* **2021**, *1*, 100012. https://doi.org/10.1016/j.crchbi.2021.100012.
(65) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684.
(66) Wei, H.; McCammon, J. A. Structure and Dynamics in Drug Discovery. *Npj Drug Discov.* **2024**, *1* (1), 1. https://doi.org/10.1038/s44386-024-00001-2.
(67) Xu, W.; Kang, C. Fragment-Based Drug Design: From Then until Now, and Toward the Future. *J. Med. Chem.* **2025**, *68* (5), 5000–5004. https://doi.org/10.1021/acs.jmedchem.5c00424.
(68) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
(69) Schaffhausen, J. Advances in Structure-Based Drug Design. *Trends Pharmacol. Sci.* **2012**, *33* (5), 223. https://doi.org/10.1016/j.tips.2012.03.011.
(70) Ferreira, F. J. N.; Carneiro, A. S. AI-Driven Drug Discovery: A Comprehensive Review. *ACS Omega* **2025**, *10* (23), 23889–23903. https://doi.org/10.1021/acsomega.5c00549.
(71) van Tilborg, D.; Brinkmann, H.; Criscuolo, E.; Rossen, L.; Özçelik, R.; Grisoni, F. Deep Learning for Low-Data Drug Discovery: Hurdles and Opportunities. *Curr. Opin. Struct. Biol.* **2024**, *86*, 102818. https://doi.org/10.1016/j.sbi.2024.102818.

(72) Isert, C.; Atz, K.; Schneider, G. Structure-Based Drug Design with Geometric Deep Learning. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102548. https://doi.org/10.1016/j.sbi.2023.102548.
(73) McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R. GNINA 1.0: Molecular Docking with Deep Learning. *J. Cheminformatics* **2021**, *13* (1), 43. https://doi.org/10.1186/s13321-021-00522-2.
(74) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58* (2), 287–296. https://doi.org/10.1021/acs.jcim.7b00650.
(75) Zeng, X.; Wang, F.; Luo, Y.; Kang, S.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep Generative Molecular Design Reshapes Drug Discovery. *Cell Rep. Med.* **2022**, *3* (12), 100794. https://doi.org/10.1016/j.xcrm.2022.100794.
(76) Green, H.; R. Koes, D.; D. Durrant, J. DeepFrag: A Deep Convolutional Neural Network for Fragment-Based Lead Optimization. *Chem. Sci.* **2021**, *12* (23), 8036–8047. https://doi.org/10.1039/D1SC00163A.
(77) Powers, A. S.; Yu, H. H.; Suriana, P.; Koodli, R. V.; Lu, T.; Paggi, J. M.; Dror, R. O. Geometric Deep Learning for Structure-Based Ligand Design. *ACS Cent. Sci.* **2023**, *9* (12), 2257–2267. https://doi.org/10.1021/acscentsci.3c00572.
(78) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9* (1), 71. https://doi.org/10.1186/1741-7007-9-71.
(79) Leach, A. *Molecular Modelling: Principles and Applications*; Harlow, England ; New York, 2001.
(80) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12%253C1463::AID-JCC4%253E3.0.CO;2-H.
(81) Braun, E.; Gilmer, J.; Mayes, H. B.; Mobley, D. L.; Monroe, J. I.; Prasad, S.; Zuckerman, D. M. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2019**, *1* (1), 5957–5957. https://doi.org/10.33011/livecoms.1.1.5957.
(82) Lopes, P. E. M.; Guvench, O.; MacKerell, A. D. Current Status of Protein Force Fields for Molecular Dynamics. *Methods Mol. Biol. Clifton NJ* **2015**, *1215*, 47–71. https://doi.org/10.1007/978-1-4939-1465-4_3.
(83) Mackerell Jr., A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25* (13), 1584–1604. https://doi.org/10.1002/jcc.20082.
(84) Chipot, C. Recent Advances in Simulation Software and Force Fields: Their Importance in Theoretical and Computational Chemistry and Biophysics. *J. Phys. Chem. B* **2024**, *128* (49), 12023–12026. https://doi.org/10.1021/acs.jpcb.4c06231.
(85) Lin, F.-Y.; MacKerell, A. D. Force Fields for Small Molecules. *Methods Mol. Biol. Clifton NJ* **2019**, *2022*, 21–54. https://doi.org/10.1007/978-1-4939-9608-7_2.
(86) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.
(87) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in

Solution. *J. Chem. Theory Comput.* **2020**, *16* (1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591.
(88) Love, O.; Galindo-Murillo, R.; Zgarbová, M.; Šponer, J.; Jurečka, P.; Cheatham, T. E. I. Assessing the Current State of Amber Force Field Modifications for DNA─2023 Edition. *J. Chem. Theory Comput.* **2023**, *19* (13), 4299–4307. https://doi.org/10.1021/acs.jctc.3c00233.
(89) Mlýnský, V.; Kührová, P.; Stadlbauer, P.; Krepl, M.; Otyepka, M.; Banáš, P.; Šponer, J. Simple Adjustment of Intranucleotide Base-Phosphate Interaction in the OL3 AMBER Force Field Improves RNA Simulations. *J. Chem. Theory Comput.* **2023**, *19* (22), 8423–8433. https://doi.org/10.1021/acs.jctc.3c00990.
(90) Dickson, C. J.; Walker, R. C.; Gould, I. R. Lipid21: Complex Lipid Membrane Simulations with AMBER. *J. Chem. Theory Comput.* **2022**, *18* (3), 1726–1736. https://doi.org/10.1021/acs.jctc.1c01217.
(91) Nagarajan, B.; Sankaranarayanan, N. V.; Desai, U. R. Perspective on Computational Simulations of Glycosaminoglycans. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2019**, *9* (2), e1388. https://doi.org/10.1002/wcms.1388.
(92) Croitoru, A.; Kumar, A.; Lambry, J.-C.; Lee, J.; Sharif, S.; Yu, W.; MacKerell, A. D. Jr.; Aleksandrov, A. Increasing the Accuracy and Robustness of the CHARMM General Force Field with an Expanded Training Set. *J. Chem. Theory Comput.* **2025**, *21* (6), 3044–3065. https://doi.org/10.1021/acs.jctc.5c00046.
(93) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236. https://doi.org/10.1021/ja9621760.
(94) Muscat, S.; Martino, G.; Manigrasso, J.; Marcia, M.; De Vivo, M. On the Power and Challenges of Atomistic Molecular Dynamics to Investigate RNA Molecules. *J. Chem. Theory Comput.* **2024**, *20* (16), 6992–7008. https://doi.org/10.1021/acs.jctc.4c00773.
(95) Tan, D.; Piana, S.; Dirks, R. M.; Shaw, D. E. RNA Force Field with Accuracy Comparable to State-of-the-Art Protein Force Fields. *Proc. Natl. Acad. Sci.* **2018**, *115* (7), E1346–E1355. https://doi.org/10.1073/pnas.1713027115.
(96) Duignan, T. T. The Potential of Neural Network Potentials. *ACS Phys. Chem. Au* **2024**, *4* (3), 232–241. https://doi.org/10.1021/acsphyschemau.4c00004.
(97) Kocer, E.; Ko, T. W.; Behler, J. Neural Network Potentials: A Concise Overview of Methods. *Annu. Rev. Phys. Chem.* **2022**, *73* (Volume 73, 2022), 163–186. https://doi.org/10.1146/annurev-physchem-082720-034254.
(98) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), 146401. https://doi.org/10.1103/PhysRevLett.98.146401.
(99) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192–3203. https://doi.org/10.1039/C6SC05720A.
(100) Simeon, G.; Fabritiis, G. de. TensorNet: Cartesian Tensor Representations for Efficient Learning of Molecular Potentials. arXiv October 30, 2023. https://doi.org/10.48550/arXiv.2306.06482.
(101) Simeon, G.; Mirarchi, A.; Pelaez, R. P.; Galvelis, R.; De Fabritiis, G. Broadening the Scope of Neural Network Potentials through Direct Inclusion of Additional Molecular

Attributes. *J. Chem. Theory Comput.* **2025**, *21* (4), 1831–1837. https://doi.org/10.1021/acs.jctc.4c01625.
(102) Chipot, C. Frontiers in Free-Energy Calculations of Biological Systems. *WIREs Comput. Mol. Sci.* **2014**, *4* (1), 71–89. https://doi.org/10.1002/wcms.1157.
(103) York, D. M. Modern Alchemical Free Energy Methods for Drug Discovery Explained. *ACS Phys. Chem. Au* **2023**, *3* (6), 478–491. https://doi.org/10.1021/acsphyschemau.3c00033.
(104) Lee, T.-S.; Lin, Z.; Allen, B. K.; Lin, C.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Sherman, W.; York, D. M. Improved Alchemical Free Energy Calculations with Optimized Smoothstep Softcore Potentials. *J. Chem. Theory Comput.* **2020**. https://doi.org/10.1021/acs.jctc.0c00237.
(105) Mey, A. S. J. S.; Allen, B.; Macdonald, H. E. B.; Chodera, J. D.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations. *Living J. Comput. Mol. Sci.* **2020**, *2* (1). https://doi.org/10.33011/livecoms.2.1.18378.
(106) Wu, J. Z.; Azimi, S.; Khuttan, S.; Deng, N.; Gallicchio, E. Alchemical Transfer Approach to Absolute Binding Free Energy Estimation. *J. Chem. Theory Comput.* **2021**, *17* (6), 3309–3319. https://doi.org/10.1021/acs.jctc.1c00266.
(107) Azimi, S.; Khuttan, S.; Wu, J. Z.; Pal, R. K.; Gallicchio, E. Relative Binding Free Energy Calculations for Ligands with Diverse Scaffolds with the Alchemical Transfer Method. *J. Chem. Inf. Model.* **2022**, *62* (2), 309–323. https://doi.org/10.1021/acs.jcim.1c01129.
(108) Sabanés Zariquiey, F.; Pérez, A.; Majewski, M.; Gallicchio, E.; De Fabritiis, G. Validation of the Alchemical Transfer Method for the Estimation of Relative Binding Affinities of Molecular Series. *J. Chem. Inf. Model.* **2023**, *63* (8), 2438–2444. https://doi.org/10.1021/acs.jcim.3c00178.
(109) Zhang, B. W.; Arasteh, S.; Levy, R. M. The UWHAM and SWHAM Software Package. *Sci. Rep.* **2019**, *9* (1), 2803. https://doi.org/10.1038/s41598-019-39420-x.
(110) Sabanés Zariquiey, F.; Farr, S. E.; Doerr, S.; De Fabritiis, G. QuantumBind-RBFE: Accurate Relative Binding Free Energy Calculations Using Neural Network Potentials. *J. Chem. Inf. Model.* **2025**, *65* (8), 4081–4089. https://doi.org/10.1021/acs.jcim.5c00033.
(111) Sohil, F.; Sohali, M. U.; Shabbir, J. An Introduction to Statistical Learning with Applications in R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. *Stat. Theory Relat. Fields* **2022**, *6* (1), 87–87. https://doi.org/10.1080/24754269.2021.1980261.
(112) Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, Second edition.; O'Reilly Media, Inc: Beijing [China] ; Sebastopol, CA, 2019.
(113) Gao, W.; Mahajan, S. P.; Sulam, J.; Gray, J. J. Deep Learning in Protein Structural Modeling and Design. *Patterns* **2020**, *1* (9), 100142. https://doi.org/10.1016/j.patter.2020.100142.
(114) Vasilev, I.; Slater, D.; Spacagna, G.; Roelants, P.; Zocca, V. *Python Deep Learning: Exploring Deep Learning Techniques and Neural Network Architectures with PyTorch, Keras, and TensorFlow*, Second edition.; Packt Publishing Limited: Birmingham Mumbai, 2019.

(115) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; Cambridge, Massachusetts, 2016.
(116) Freudenthal, B. D.; Beard, W. A.; Shock, D. D.; Wilson, S. H. Observing a DNA Polymerase Choose Right from Wrong. *Cell* **2013**, *154* (1), 157–168. https://doi.org/10.1016/j.cell.2013.05.048.
(117) Yang, L.; Beard, W. A.; Wilson, S. H.; Broyde, S.; Schlick, T. Highly Organized but Pliant Active Site of DNA Polymerase β: Compensatory Mechanisms in Mutant Enzymes Revealed by Dynamics Simulations and Energy Analyses. *Biophys. J.* **2004**, *86* (6), 3392–3408. https://doi.org/10.1529/biophysj.103.036012.
(118) Genna, V.; Colombo, M.; De Vivo, M.; Marcia, M. Second-Shell Basic Residues Expand the Two-Metal-Ion Architecture of DNA and RNA Processing Enzymes. *Structure* **2018**, *26* (1), 40-50.e2. https://doi.org/10.1016/j.str.2017.11.008.
(119) Li, Y.; Gridley, C. L.; Jaeger, J.; Sweasy, J. B.; Schlick, T. Unfavorable Electrostatic and Steric Interactions in DNA Polymerase β E295K Mutant Interfere with the Enzyme's Pathway. *J. Am. Chem. Soc.* **2012**, *134* (24), 9999–10010. https://doi.org/10.1021/ja300361r.
(120) Batra, V. K.; Perera, L.; Lin, P.; Shock, D. D.; Beard, W. A.; Pedersen, L. C.; Pedersen, L. G.; Wilson, S. H. Amino Acid Substitution in the Active Site of DNA Polymerase β Explains the Energy Barrier of the Nucleotidyl Transfer Reaction. *J. Am. Chem. Soc.* **2013**, *135* (21), 8078–8088. https://doi.org/10.1021/ja403842j.
(121) Kirby, T. W.; DeRose, E. F.; Cavanaugh, N. A.; Beard, W. A.; Shock, D. D.; Mueller, G. A.; Wilson, S. H.; London, R. E. Metal-Induced DNA Translocation Leads to DNA Polymerase Conformational Activation. *Nucleic Acids Res.* **2012**, *40* (7), 2974–2983. https://doi.org/10.1093/nar/gkr1218.
(122) Kraynov, V. S.; Werneburg, B. G.; Zhong, X.; Lee, H.; Ahn, J.; Tsai, M. D. DNA Polymerase Beta: Analysis of the Contributions of Tyrosine-271 and Asparagine-279 to Substrate Specificity and Fidelity of DNA Replication by Pre-Steady-State Kinetics. *Biochem. J.* **1997**, *323* (Pt 1), 103–111.
(123) Yariv, B.; Yariv, E.; Kessel, A.; Masrati, G.; Chorin, A. B.; Martz, E.; Mayrose, I.; Pupko, T.; Ben-Tal, N. Using Evolutionary Data to Make Sense of Macromolecules with a "Face-Lifted" ConSurf. *Protein Sci.* **2023**, *32* (3), e4582. https://doi.org/10.1002/pro.4582.
(124) Cavanaugh, N. A.; Beard, W. A.; Batra, V. K.; Perera, L.; Pedersen, L. G.; Wilson, S. H. Molecular Insights into DNA Polymerase Deterrents for Ribonucleotide Insertion. *J. Biol. Chem.* **2011**, *286* (36), 31650–31660. https://doi.org/10.1074/jbc.M111.253401.
(125) Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Liles, K.; Chun, M.; Li, P.; Gohara, D. W.; Dolinsky, T.; Konecny, R.; Koes, D. R.; Nielsen, J. E.; Head-Gordon, T.; Geng, W.; Krasny, R.; Wei, G.-W.; Holst, M. J.; McCammon, J. A.; Baker, N. A. Improvements to the APBS Biomolecular Solvation Software Suite. *Protein Sci.* **2018**, *27* (1), 112–128. https://doi.org/10.1002/pro.3280.
(126) Li, S.; Olson, W. K.; Lu, X.-J. Web 3DNA 2.0 for the Analysis, Visualization, and Modeling of 3D Nucleic Acid Structures. *Nucleic Acids Res.* **2019**, *47* (W1), W26–W34. https://doi.org/10.1093/nar/gkz394.
(127) Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8, 2015.
(128) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpí, J. L.; González, C.;

Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. Parmbsc1: A Refined Force Field for DNA Simulations. *Nat. Methods* **2016**, *13* (1), 55–58. https://doi.org/10.1038/nmeth.3658.
(129) Allnér, O.; Nilsson, L.; Villa, A. Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **2012**, *8* (4), 1493–1502. https://doi.org/10.1021/ct3000734.
(130) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* **2010**, *12* (28), 7821–7839. https://doi.org/10.1039/C0CP00111B.
(131) Meng, Y.; Sabri Dashti, D.; Roitberg, A. E. Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations. *J. Chem. Theory Comput.* **2011**, *7* (9), 2721–2727. https://doi.org/10.1021/ct200153u.
(132) Maximoff, S. N.; Kamerlin, S. C. L.; Florián, J. DNA Polymerase λ Active Site Favors a Mutagenic Mispair between the Enol Form of Deoxyguanosine Triphosphate Substrate and the Keto Form of Thymidine Template: A Free Energy Perturbation Study. *J. Phys. Chem. B* **2017**, *121* (33), 7813–7822. https://doi.org/10.1021/acs.jpcb.7b04874.
(133) Gapsys, V.; de Groot, B. L. Alchemical Free Energy Calculations for Nucleotide Mutations in Protein–DNA Complexes. *J. Chem. Theory Comput.* **2017**, *13* (12), 6275–6289. https://doi.org/10.1021/acs.jctc.7b00849.
(134) Geronimo, I.; De Vivo, M. Alchemical Free-Energy Calculations of Watson–Crick and Hoogsteen Base Pairing Interconversion in DNA. *J. Chem. Theory Comput.* **2022**, *18* (11), 6966–6973. https://doi.org/10.1021/acs.jctc.2c00848.
(135) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60* (11), 5595–5623. https://doi.org/10.1021/acs.jcim.0c00613.
(136) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. *J. Comput. Chem.* **2015**, *36* (5), 348–354. https://doi.org/10.1002/jcc.23804.
(137) Wade, A. D.; Bhati, A. P.; Wan, S.; Coveney, P. V. Alchemical Free Energy Estimators and Molecular Dynamics Engines: Accuracy, Precision, and Reproducibility. *J. Chem. Theory Comput.* **2022**, *18* (6), 3972–3987. https://doi.org/10.1021/acs.jctc.2c00114.
(138) Ross, G. A.; Lu, C.; Scarabelli, G.; Albanese, S. K.; Houang, E.; Abel, R.; Harder, E. D.; Wang, L. The Maximal and Current Accuracy of Rigorous Protein-Ligand Binding Free Energy Calculations. *Commun. Chem.* **2023**, *6* (1), 222. https://doi.org/10.1038/s42004-023-01019-9.
(139) Chen, W.; Cui, D.; Jerome, S. V.; Michino, M.; Lenselink, E. B.; Huggins, D. J.; Beautrait, A.; Vendome, J.; Abel, R.; Friesner, R. A.; Wang, L. Enhancing Hit Discovery in Virtual Screening through Absolute Protein–Ligand Binding Free-Energy Calculations. *J. Chem. Inf. Model.* **2023**, *63* (10), 3171–3185. https://doi.org/10.1021/acs.jcim.3c00013.
(140) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko,

M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137* (7), 2695–2703. https://doi.org/10.1021/ja512751q.

(141) Amezcua, M.; Setiadi, J.; Ge, Y.; Mobley, D. L. An Overview of the SAMPL8 Host–Guest Binding Challenge. *J. Comput. Aided Mol. Des.* **2022**, *36* (10), 707–734. https://doi.org/10.1007/s10822-022-00462-5.

(142) Parks, C. D.; Gaieb, Z.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Jansen, J. M.; McGaughey, G.; Lewis, R. A.; Bembenek, S. D.; Ameriks, M. K.; Mirzadegan, T.; Burley, S. K.; Amaro, R. E.; Gilson, M. K. D3R Grand Challenge 4: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided Mol. Des.* **2020**, *34* (2), 99–119. https://doi.org/10.1007/s10822-020-00289-y.

(143) Yang, G.; Franklin, M.; Li, J.; Lin, T.-C.; Konigsberg, W. A Conserved Tyr Residue Is Required for Sugar Selectivity in a Pol α DNA Polymerase. *Biochemistry* **2002**, *41* (32), 10256–10261. https://doi.org/10.1021/bi0202171.

(144) Lasken, R. S.; Goodman, M. F. A Fidelity Assay Using "Dideoxy" DNA Sequencing: A Measurement of Sequence Dependence and Frequency of Forming 5-Bromouracil X Guanine Base Mispairs. *Proc. Natl. Acad. Sci.* **1985**, *82* (5), 1301–1305. https://doi.org/10.1073/pnas.82.5.1301.

(145) Kufe, D.; Spriggs, D.; Egan, E. M.; Munroe, D. Relationships Among Ara-CTP Pools, Formation of (Ara-C)DNA, and Cytotoxicity of Human Leukemic Cells. *Blood* **1984**, *64* (1), 54–58. https://doi.org/10.1182/blood.V64.1.54.54.

(146) Joung, I. S.; Cheatham, T. E. I. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(147) Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Cisneros, G. A.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham, T. E. I.; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz, K. M. Jr. AmberTools. *J. Chem. Inf. Model.* **2023**, *63* (20), 6183–6191. https://doi.org/10.1021/acs.jcim.3c01153.

(148) Özçelik, R.; Brinkmann, H.; Criscuolo, E.; Grisoni, F. Generative Deep Learning for de Novo Drug Design─A Chemical Space Odyssey. *J. Chem. Inf. Model.* **2025**, *65* (14), 7352–7372. https://doi.org/10.1021/acs.jcim.5c00641.

(149) Zhang, K.; Yang, X.; Wang, Y.; Yu, Y.; Huang, N.; Li, G.; Li, X.; Wu, J. C.; Yang, S. Artificial Intelligence in Drug Development. *Nat. Med.* **2025**, *31* (1), 45–59. https://doi.org/10.1038/s41591-024-03434-4.

(150) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50* (2), 302–309. https://doi.org/10.1021/acs.accounts.6b00491.

(151) Cusick, M. E.; Yu, H.; Smolyar, A.; Venkatesan, K.; Carvunis, A.-R.; Simonis, N.; Rual, J.-F.; Borick, H.; Braun, P.; Dreze, M.; Vandenhaute, J.; Galli, M.; Yazaki, J.; Hill, D. E.;

Ecker, J. R.; Roth, F. P.; Vidal, M. Literature-Curated Protein Interaction Datasets. *Nat. Methods* **2009**, *6* (1), 39–46. https://doi.org/10.1038/nmeth.1284.
(152) Moine-Franel, A.; Mareuil, F.; Nilges, M.; Ciambur, C. B.; Sperandio, O. A Comprehensive Dataset of Protein-Protein Interactions and Ligand Binding Pockets for Advancing Drug Discovery. *Sci. Data* **2024**, *11* (1), 402. https://doi.org/10.1038/s41597-024-03233-z.
(153) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinforma. Oxf. Engl.* **2017**, *33* (19), 3036–3042. https://doi.org/10.1093/bioinformatics/btx350.
(154) Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; Koes, D. R. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *J. Chem. Inf. Model.* **2020**, *60* (9), 4200–4215. https://doi.org/10.1021/acs.jcim.0c00411.
(155) Besharatifard, M.; Vafaee, F. A Review on Graph Neural Networks for Predicting Synergistic Drug Combinations. *Artif. Intell. Rev.* **2024**, *57* (3), 49. https://doi.org/10.1007/s10462-023-10669-z.
(156) Zhang, Z.; Chen, L.; Zhong, F.; Wang, D.; Jiang, J.; Zhang, S.; Jiang, H.; Zheng, M.; Li, X. Graph Neural Network Approaches for Drug-Target Interactions. *Curr. Opin. Struct. Biol.* **2022**, *73*, 102327. https://doi.org/10.1016/j.sbi.2021.102327.
(157) Lai, H.; Wang, L.; Qian, R.; Huang, J.; Zhou, P.; Ye, G.; Wu, F.; Wu, F.; Zeng, X.; Liu, W. Interformer: An Interaction-Aware Model for Protein-Ligand Docking and Affinity Prediction. *Nat. Commun.* **2024**, *15* (1), 10223. https://doi.org/10.1038/s41467-024-54440-6.
(158) Zhang, S.; Huo, D.; Horne, R. I.; Qi, Y.; Pujalte Ojeda, S.; Yan, A.; Vendruscolo, M. Sequence-Based Virtual Screening Using Transformers. *Nat. Commun.* **2025**, *16* (1), 6925. https://doi.org/10.1038/s41467-025-61833-8.
(159) Zhang, Z.; Yan, J.; Huang, Y.; Liu, Q.; Chen, E.; Wang, M.; Zitnik, M. Geometric Deep Learning for Structure-Based Drug Design: A Survey. arXiv November 16, 2024. https://doi.org/10.48550/arXiv.2306.11768.
(160) Bai, Q.; Xu, T.; Huang, J.; Pérez-Sánchez, H. Geometric Deep Learning Methods and Applications in 3D Structure-Based Drug Design. *Drug Discov. Today* **2024**, *29* (7), 104024. https://doi.org/10.1016/j.drudis.2024.104024.
(161) Zeng, X.; Wang, F.; Luo, Y.; Kang, S.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep Generative Molecular Design Reshapes Drug Discovery. *Cell Rep. Med.* **2022**, *3* (12). https://doi.org/10.1016/j.xcrm.2022.100794.
(162) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), e1608. https://doi.org/10.1002/wcms.1608.
(163) Kanakala, G. C.; Devata, S.; Chatterjee, P.; Priyakumar, U. D. Generative Artificial Intelligence for Small Molecule Drug Design. *Curr. Opin. Biotechnol.* **2024**, *89*, 103175. https://doi.org/10.1016/j.copbio.2024.103175.
(164) Rose, Y.; Duarte, J. M.; Lowe, R.; Segura, J.; Bi, C.; Bhikadiya, C.; Chen, L.; Rose, A. S.; Bittrich, S.; Burley, S. K.; Westbrook, J. D. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. *J. Mol. Biol.* **2021**, *433* (11), 166704. https://doi.org/10.1016/j.jmb.2020.11.003.

(165) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8. https://doi.org/10.1186/1758-2946-1-8.
(166) Kruger, F.; Stiefl, N.; Landrum, G. A. rdScaffoldNetwork: The Scaffold Network Implementation in RDKit. *J. Chem. Inf. Model.* **2020**, *60* (7), 3331–3335. https://doi.org/10.1021/acs.jcim.0c00296.
(167) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35* (11), 1026–1028. https://doi.org/10.1038/nbt.3988.
(168) Unoh, Y.; Uehara, S.; Nakahara, K.; Nobori, H.; Yamatsu, Y.; Yamamoto, S.; Maruyama, Y.; Taoda, Y.; Kasamatsu, K.; Suto, T.; Kouki, K.; Nakahashi, A.; Kawashima, S.; Sanaki, T.; Toba, S.; Uemura, K.; Mizutare, T.; Ando, S.; Sasaki, M.; Orba, Y.; Sawa, H.; Sato, A.; Sato, T.; Kato, T.; Tachibana, Y. Discovery of S-217622, a Noncovalent Oral SARS-CoV-2 3CL Protease Inhibitor Clinical Candidate for Treating COVID-19. *J. Med. Chem.* **2022**, *65* (9), 6499–6512. https://doi.org/10.1021/acs.jmedchem.2c00117.
(169) Ward, R. A.; Bethel, P.; Cook, C.; Davies, E.; Debreczeni, J. E.; Fairley, G.; Feron, L.; Flemington, V.; Graham, M. A.; Greenwood, R.; Griffin, N.; Hanson, L.; Hopcroft, P.; Howard, T. D.; Hudson, J.; James, M.; Jones, C. D.; Jones, C. R.; Lamont, S.; Lewis, R.; Lindsay, N.; Roberts, K.; Simpson, I.; St-Gallay, S.; Swallow, S.; Tang, J.; Tonge, M.; Wang, Z.; Zhai, B. Structure-Guided Discovery of Potent and Selective Inhibitors of ERK1/2 from a Modestly Active and Promiscuous Chemical Start Point. *J. Med. Chem.* **2017**, *60* (8), 3438–3450. https://doi.org/10.1021/acs.jmedchem.7b00267.
(170) Schoenfeld, R. C.; Bourdet, D. L.; Brameld, K. A.; Chin, E.; de Vicente, J.; Fung, A.; Harris, S. F.; Lee, E. K.; Le Pogam, S.; Leveque, V.; Li, J.; Lui, A. S.-T.; Najera, I.; Rajyaguru, S.; Sangi, M.; Steiner, S.; Talamas, F. X.; Taygerly, J. P.; Zhao, J. Discovery of a Novel Series of Potent Non-Nucleoside Inhibitors of Hepatitis C Virus NS5B. *J. Med. Chem.* **2013**, *56* (20), 8163–8182. https://doi.org/10.1021/jm401266k.

Activities PhD

David Ricardo Figueroa Blanco.

During the period 2021/2022. I attend the following courses of the PhD in data science and computation.

- Computational Methods for Life Sciences, a cura di IIT, (Docenti: IIT); Numero crediti 7
- Supervised Statistical Learning (Docente: Prof. Laura Anderlucci) Numero crediti: 6

During period 2022/2023. I attend the following courses:

- Deep Learning (Docente: Prof. Andrea Asperti) Numero crediti: 6 – Sede UNIBO
- Software & Computing for Applied Physics(Docente: Prof. Enrico Giampieri) NumeroCrediti: 6

Other activities to full fill the number of free credits were:

- 23rd Bologna Winter School "Structural Bioinformatics in the era of AlphaFold2". February 9-25, 2022. 18 Hours. 2 CFU.
- Multiscale Molecular Dynamics with MiMiC July 18, 2022 - July 22, 2022 25 Hours. 3 CFU.
- Advanced school in Drug Research and Development Integrating Structural and Biophysical Data in Drug Discovery in the Artificial Intelligence Era. STARTS September 4-9. 3 CFU.
- SBDD 2023 EDITION. Computational Advances in Drug Discovery May 2-4, 2023 Sestri Levante, Italy. 3 CFU
- Advance Learning Algorithms 31 hours. Coursera-Stanford University. 3 CFU.
- Unsupervised Learning, Recommenders, Reinforcement Learning. 26 hours 3 CFU.

Additionally, I attended and presented my work in the conferences:

The 10th Annual CCPBioSim & MGMS 2024 Conference: Molecular modelling in structure-based drug design. Newcasltle UK. July 1-3. 2024.

Structure-Based Drug Design (SBDD) Conference 2025. COMPUTATIONAL ADVANCES IN DRUG DESIGN. Sestri Levante. October 1-3. 2025

Publication list.

Figueroa Blanco, D. R.; Vidossich, P.; De Vivo, M. Correct Nucleotide Selection Is Confined at the Binding Site of Polymerase Enzymes. J. Chem. Inf. Model. 2024, 64 (13), 5285–5294. https://doi.org/10.1021/acs.jcim.4c00696.